

QUANTUM FIELD THEORY: FOCK-BORN FRAMEWORK

LATTICE-FIRST VIEW

FOCK LAYER

THE GAME ENGINE

Flat lattice. Unitary evolution.
All possibilities.

$$n_1, n_2, \dots, n_N$$

$$\dot{h}_t = -iHh_t$$

$$h_t$$

Fock space $\mathcal{F}$

Hidden state $h_t \in \mathcal{F}$

BORN LAYER

THE VR INTERFACE

Curved spacetime. Classical outcomes.
What the observer experiences.

Outcome λ

Probability $P(\lambda)$

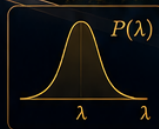
$$P(\lambda) = \langle h_t | P_\lambda | h_t \rangle$$

P_λ = projector onto
eigspace E_λ

$$\hat{O}$$

Observable
(choice of
resolution)

Mass = M
Energy = E
Spin = s
...



FLAT LATTICE

Discrete, combinatorial
quantum substrate



$$\hat{O}$$

Observer chooses
what can be resolved



CURVED SPACETIME

Emergent geometry.
Classical world.

THE UNIVERSE COMPUTES IN FOCK SPACE.
OBSERVATION RENDERS A WORLD.

YING NIAN WU

UCLA DEPARTMENT OF
STATISTICS AND DATA SCIENCE

Quantum Field Theory

The Fock-Born Framework and the Lattice-First View

Ying Nian Wu
UCLA Department of Statistics and Data Science

Version June 2026, written with the help of Claude 4.6 (book cover by GPT 5.5)

Contents

1	Preface: The Fock-Born Framework and the Lattice-First View	14
2	Dirac Notation, Basis Agnosticism, Direct Product, and Direct Sum	29
2.1	The problem with dummy names	29
2.2	Kets, bras, and the inner product	29
2.3	Dirac notation as matrix and vector notation	30
2.4	One-hot vectors, dense vectors, and the continuous index	32
2.5	The wavefunction is a coefficient, not a function	34
2.6	Word embedding and position embedding in transformers	36
2.7	The power of Dirac notation: basis agnosticism	37
2.8	Fock basis vectors as embeddings of world states	40
2.9	Superposition: the hidden state as a weighted sum of basis vectors	41
2.10	Making Fock space concrete: direct product, direct sum, and the full construction	43
3	Quantum Mechanics as an RNN: The Core Model	51
3.1	The Fock-Born architecture	51
3.2	The Fock layer: a single vector, rotating	53
3.3	The Hamiltonian: the motor of the game engine	54
3.4	The Born layer: a multi-lingual interactive output layer	56
3.5	The observer as a primitive	59
3.6	A concrete example: the qubit	60
4	Hamiltonians, Quantization, and Observables	63
4.1	The classical Hamiltonian: total energy	63
4.2	Quantization: promoting scalars to operators	63
4.3	The canonical commutation relation	64
4.4	Observables as the observer's choice: vocabulary, probing vectors, and degeneracy	65
4.5	Energy as observable: the Hamiltonian's dual role	67
4.6	Observables are Hermitian: eigenvalues are real	69
4.7	The uncertainty principle as incompatible tokenisers	70
5	The Schrödinger equation: three formulations and one theory	72
5.1	Heisenberg's matrix mechanics (1925): the commutator encodes the dynamics .	72
5.2	Schrödinger's wave mechanics (1926): the wave equation	73
5.3	Dirac's q-numbers (1925–26): the logical leap	74
5.4	Dirac's transformation theory (1927): the unification	75

5.5	Dirac's Lagrangian paper (1933): the seed of the path integral	77
5.6	Three formulations, one vector, many lenses	79
6	The Harmonic Oscillator: A Complete Treatment	81
6.1	The Hamiltonian, ladder operators, and the spectrum	81
6.2	The ground state and excited states: explicit wavefunctions	84
6.3	Concrete matrix representations	85
6.4	The Hamiltonian motor: free and interacting	86
6.5	The vacuum: one quiet vector	86
6.6	The catalog of observables: what the Born layer can measure	88
6.7	Perturbation theory and the Dyson series	90
6.8	Coherent states and the classical limit	93
6.9	Condensates and spontaneous symmetry breaking	95
6.10	Squeezed states	95
6.11	The harmonic oscillator as the atom of quantum field theory	96
7	The Feynman Path Integral	98
7.1	The propagator as a Fock-layer matrix element	98
7.2	Time slicing: matrix multiplication, not history	98
7.3	The short-time matrix element and the Lagrangian	99
7.4	The path integral formula	100
7.5	Stationary phase and the classical trajectory	101
7.6	The harmonic oscillator path integral	101
7.7	Perturbation theory, Wick's theorem, and Feynman diagrams	103
7.8	The generating function and the bridge to Born-layer observables	106
7.9	Path integral and Markov chain: amplitude versus probability	109
8	Wick Rotation: From Oscillation to Sampling	110
8.1	The substitution: $t \rightarrow -i\tau$	110
8.2	The harmonic oscillator in both times	110
8.3	The Euclidean path integral: convergent and positive	111
8.4	The interacting case: Euclidean perturbation theory	112
8.5	The Euclidean path integral as classical statistical mechanics	113
8.6	Quantum Monte Carlo: sampling the Fock layer	114
8.7	The sign problem: when Monte Carlo fails	115
8.8	Summary: the Fock-Born perspective on Wick rotation	115
8.9	Wick rotation in the Fock-Born framework	116
9	Degenerate Observables and the Projected Superposition	119
9.1	The identity observable: the starting point	119
9.2	Partial lifting of degeneracy: a concrete example	119
9.3	The general framework	120
9.4	The roots of degeneracy: why classical observables are degenerate	121
9.5	The Born layer certifies: delta or not	122
9.6	Foreshadowing	124
10	Tensor Product, Entanglement, and Decoherence	125
10.1	The tensor product: the joint hidden state	125
10.2	The identity on B : the most important source of degeneracy	125
10.3	Measuring A alone: the Born-Bohr rule applied	126
10.4	Entanglement: when the coefficients do not factorise	126
10.5	Bell's theorem: classical values do not exist between measurements	127

10.6	Spooky action is not spooky: it is just $\hbar$	127
10.7	The density matrix from the degeneracy framework	128
10.8	Pure states and mixed states	129
10.9	Temperature from the tensor product degeneracy	129
10.10	Decoherence: the environment provides enormous degeneracy	130
10.11	The degeneracy hierarchy, unified	131
11	Angular Momentum: the First Representation	133
11.1	The angular momentum algebra	133
11.2	The eigenvalues: ladder operators	133
11.3	The algebra is $\mathfrak{su}(2) \cong \mathfrak{so}(3)$	134
11.4	Spherical harmonics: the position-basis representation	135
11.5	Angular momentum as an observable: the Born rule	136
11.6	What this foreshadows	137
11.7	The two-valuedness: orbital angular momentum is not enough	137
12	Spin and the Pauli Equation	141
12.1	Enriched basis vectors: the spinor index	141
12.2	The Hamiltonian acts on both indices	142
12.3	Pauli matrices: the spin operators	142
12.4	The Pauli Hamiltonian motor	143
12.5	Direct product operators acting on direct product kets	144
12.6	Total angular momentum: spin is not a spinning top	145
12.7	The hidden state as an $\infty \times 2$ matrix: two-sided operator action	147
13	The Dirac Equation: Spin from Spacetime	149
13.1	The Dirac equation: why nature keeps it simple	149
13.2	Special relativity: the coupling of time and space	151
13.3	The relativistic wave equation: Klein-Gordon and its failure	154
13.4	Dirac's discovery: linearizing the relativistic Hamiltonian	155
13.5	The Dirac Hamiltonian motor	157
13.6	Why physicists call it the most beautiful equation	161
13.7	Spin emerges automatically: comparing Dirac and Pauli	162
13.8	The crisis: negative energy states	163
13.9	Dirac's hole theory: the first glimpse of the vacuum	164
13.10	The Fock space: abandoning single-particle quantum mechanics	165
13.11	Pauli's reconceptualization: the Dirac equation as a classical field	167
13.12	What Dirac actually discovered	168
14	Quantizing the Electromagnetic Field	172
14.1	An irony in the history of quantum field theory	172
14.2	Classical electromagnetism in tensor language	172
14.3	The electromagnetic Hamiltonian motor	173
14.4	The classical EM field as coupled oscillators	173
14.5	Quantization: each mode becomes a quantum oscillator	173
14.6	The Fock space of photons	174
14.7	The Born rule for photons: projected superpositions in Fock space	174
14.8	Coherent states: the bridge to classical EM	174
14.9	The emergence of the classical EM field	175
14.10	The photon vacuum	175
14.11	The RNN picture of the quantized EM field	175
14.12	Dirac's original construction: from action-angle variables to creation operators	176

15 Many-Particle Quantum Mechanics: Fixed Particle Number	177
15.1 The naive n -particle space and its problem	177
15.2 Bosons and fermions: the exchange symmetry	177
15.3 The physical n -particle Hilbert spaces	178
15.4 Basis vectors for n identical particles	178
15.5 The Pauli exclusion principle	179
15.6 The n -particle wavefunction as a projection	180
15.7 The Hamiltonian motor for n identical particles	181
15.8 Foreshadowing Fock space: connecting particle-number sectors	186
15.9 Preview: symmetrisation and the origin of quantum statistics	188
16 Non-Relativistic Quantum Field Theory: Varying Particle Number	191
16.1 The basis vectors of Fock space	191
16.2 The Fock hidden state as an explicit superposition	191
16.3 Each basis vector is monosemantic: the Fock-SAE correspondence	192
16.4 Why this superposition belongs to Fock space: the direct sum	192
16.5 Why momentum rather than position: the natural basis	193
16.6 The Schrödinger equation on Fock space	193
16.7 Creation and annihilation operators	193
16.8 The field operator and mode expansion	194
16.9 The Hamiltonian motor as a polynomial of field operators	195
16.10 Second quantization: what the name really means	200
16.11 Physical examples: the electron gas and phonons	200
16.12 Condensate and spontaneous symmetry breaking	201
16.13 The Born rendering layer: consistency across levels	203
16.14 Time slicing and the path integral in Fock language	205
16.15 The Dyson series and Feynman diagrams	206
16.16 Preview: towards relativistic quantum field theory	206
17 Lorentz Transformations, Rotations, Boosts, and Einstein notation	208
17.1 Spacetime and the invariant interval	208
17.2 Einstein index notation: upper and lower indices	209
17.3 Lorentz transformations: definition	210
17.4 Rotations	210
17.5 Boosts	210
17.6 The Lie algebra of the Lorentz group	211
17.7 Lorentz invariance: the practical test	212
17.8 Wick rotation: Lorentz as a rotated sphere	212
17.9 Tensors: the chain rule in disguise	215
17.10 Lorentz invariance from the lattice	217
17.11 Space and time: separate in the engine, packaged in the rendering	220
18 The Scalar Field: Hamiltonian, Correlators, and Schwinger's Source	224
18.1 The free scalar field on a lattice	224
18.2 Adding interactions	225
18.3 Correlators: the central objects of the Fock layer	225
18.4 Connected correlators: genuine multi-point correlations	227
18.5 Schwinger's source theory: $Z[J]$, $W[J]$, and $\Gamma[\bar{\phi}]$ from operators	227
18.6 The spectral representation: what "a particle" means	230
18.7 The interaction picture and the Dyson series	232
19 How Is a Particle Observed?	234

19.1	A detector is a macroscopic Born-layer system	234
19.2	Scattering as a Born probability	234
19.3	The LSZ reduction formula: from correlators to cross sections	235
19.4	The anatomy of $2 \rightarrow 2$ scattering	235
19.5	What is a particle? What is a resonance?	239
19.6	What experimentalists actually report	240
19.7	The catalog of Born-layer observables for the scalar field	241
20	Renormalization: The Hamiltonian Side	242
20.1	The game engine is a finite lattice	242
20.2	The cinema analogy	243
20.3	Rendering stability: the definition	243
20.4	Why do the engine parameters depend on the lattice spacing?	244
20.5	The Feshbach formula: integrating out the short-distance shell	245
20.6	The Wegner flow and the beta function	247
20.7	Renormalization of correlators	247
20.8	What renormalisability means	248
21	The Scalar Field: Path Integral, $Z[J]$, $W[J]$, $\Gamma[\bar{\phi}]$	249
21.1	From the Hamiltonian to the path integral: time slicing for a field theory	249
21.2	The path integral computes Schwinger's $Z[J]$	251
21.3	Worked example: the free-field path integral	252
21.4	$W[J]$ and $\Gamma[\bar{\phi}]$ from the path integral	253
21.5	The saddle-point (loop) expansion	254
21.6	Feynman diagrams from the perturbative expansion	255
21.7	The $i\varepsilon$ prescription	256
21.8	The bridges to Born-layer observables	257
21.9	The rendering-stability workflow: from $\theta(a)$ to $o(a/\lambda)$	258
21.10	Wick rotation on the lattice: from Lorentz to Euclidean	260
21.11	The two languages, side by side	261
21.12	Common conflations, revisited	262
22	Renormalization: The Path Integral Side	264
22.1	The CNN binary classifier as a field theory	264
22.2	The Euclidean generating functional on the lattice	265
22.3	The one-loop effective potential: a worked example	267
22.4	Renormalization conditions on $\Gamma^{(n)}$	268
22.5	The beta function from the path integral	269
22.6	The correspondence: Hamiltonian and path integral	270
22.7	Why both sides exist	271
23	Wick Rotation, Quantum Monte Carlo, and the Heat Kernel	273
23.1	Wick rotation for the scalar field	273
23.2	Euclidean correlators and their Lorentzian counterparts	273
23.3	The Euclidean shadow: a catalog of correspondences	273
23.4	The lattice Euclidean path integral	275
23.5	Quantum Monte Carlo: what it computes	275
23.6	The sign problem	276
23.7	What the Euclidean side cannot see	277
23.8	The heat kernel: the Euclidean propagator as a tool	278
23.9	The Seeley–DeWitt expansion: UV structure without diagrams	278
23.10	Deriving the Seeley–DeWitt coefficients	279

23.11	Schwinger's proper-time method	280
23.12	Beta functions from the heat kernel	281
23.13	Anomalies and index theorems	282
23.14	The heat kernel as a unifying tool	282
23.15	The Fock-Born reading	282
24	Wilson Renormalization Group and Weinberg Effective Field Theory	284
24.1	Dimensional analysis: one rule determines everything	284
24.2	From dimensional analysis to renormalizability: the lattice-first argument	286
24.3	Wilson's renormalization group: the flow of couplings	290
24.4	Hamiltonian renormalization	292
24.5	The lattice-first synthesis	294
24.6	Equivalence of the lattice-first and Wilson views	295
24.7	Emergence of operators: zero couplings that become nonzero	298
24.8	Weinberg's effective field theory programme	304
24.9	Weinberg's E/Λ vs lattice-first a/λ : a term-by-term comparison	308
24.10	UV completion, high-energy scattering, and the lattice ceiling	311
24.11	Mass parameters on the lattice: numerical stability and the hierarchy	314
25	Lie Groups, Lie Algebras, and Their Equivariant Representations	318
25.1	Why equivariant representations?	318
25.2	From e^x to matrix exponentials	318
25.3	Lie groups: continuous symmetry groups	319
25.4	Lie algebras: infinitesimal generators	320
25.5	$SU(2)$ and $SO(3)$: the same algebra, different groups	320
25.6	Representations of Lie algebras and Lie groups	322
25.7	Representations of $\mathfrak{su}(2)$: the spin- j family	322
25.8	Why half-integer spins do not represent $SO(3)$	323
25.9	Infinite-dimensional representations and function spaces	324
25.10	Connection to equivariant representation learning	325
25.11	3D vectors as 2×2 matrices: the adjoint action	328
25.12	The Schrödinger equation as a representation of time translation	331
26	Equivariant Representations of the Lorentz Group	333
26.1	Fields as equivariant representations	333
26.2	Diagram 1: the abstract transformation law	334
26.3	Diagram 2: spacetime as 2×2 matrices	335
26.4	Diagram 3: fields on matrix spacetime	336
26.5	Chirality	338
26.6	Spinors versus vectors: the sign that makes all the difference	338
26.7	General representations: (j_L, j_R) as matrix dimensions	339
26.8	Dirac spinor	339
26.9	Diagram 3 in spin-index notation	343
26.10	Indices vs. labels: what transforms and what does not	345
26.11	From equivariant fields to an invariant Lagrangian	346
26.12	Wigner's classification: what a particle IS	349
27	Adding an Internal Indices for Gauge Transformation	354
27.1	Internal indices and local gauge invariance	354
27.2	Adding one internal index	355
27.3	The field as a 3-way tensor	356
27.4	Singlets, N -lets, and suppressing the internal index	357

27.5	Suppressing spinor indices without collapsing them	357
27.6	Tensors in ML vs. tensors in physics	358
27.7	A note on terminology: “quantum number”	359
27.8	Labels: chirality, flavour, and the actor’s name	360
27.9	The Dirac spinor in full index notation	361
27.10	Local gauge invariance $SU(N)$: matter demands the force	363
28	Multiple Internal Indices for Gauge Transformation	366
28.1	Multiple internal indices: product gauge groups	366
28.2	Adding a second internal index	366
28.3	Both internal indices acted on simultaneously	367
28.4	Full index notation with all four indices	367
28.5	Suppressing indices	367
28.6	Adding labels and the Dirac spinor	368
28.7	The Standard Model gauge group $U(1)_Y \times SU(2)_L \times SU(3)_c$	369
28.8	The Beijing Opera analogy	370
28.9	Nature is left-handed: isospin as a chiral index	371
28.10	Nature prefers minimalism over symmetry	373
29	Fields, Lagrangian, and Noether Currents	378
29.1	Three kinds of field: a reminder	378
29.2	Designing a Lagrangian: the rules of the game	378
29.3	Kinetic and potential terms	379
29.4	Noether’s theorem: symmetry dictates conservation	380
29.5	Noether currents beyond the free field: interactions, anomalies, and condensates	383
29.6	The Wald identity and masslessness protection	389
29.7	Discrete symmetries: parity, pseudoscalars, and pseudovectors	392
30	The Dance between Matter and Force: Local Gauge Invariance	397
30.1	Force fields as messengers for dimension reduction	397
30.2	Local gauge invariance and its price	398
30.3	Non-Abelian gauge invariance: $SU(N)$	400
30.4	The three gauge groups of the Standard Model	401
30.5	The adjoint representation: gauge bosons as Lie algebra elements	402
30.6	Gauge orbits: physical states as equivalence classes	405
30.7	Gauge invariance as dimension reduction	406
30.8	The messenger picture: what gauge bosons actually carry	407
30.9	Yang and Mills (1954): from $U(1)$ to $SU(2)$	409
30.10	Connection and curvature: what the gauge field is	412
31	The Higgs Mechanism: Smuggling Mass with Scalar Mole	416
31.1	What is mass in quantum field theory?	416
31.2	Why mass is forbidden for spin- $\frac{1}{2}$ and spin-1, but allowed for spin-0	418
31.3	Mass in QCD alone: quarks yes, gluons no	419
31.4	The loophole: imaginary mass for a scalar mole	420
31.5	Goldstone’s theorem and the fate of flat directions	420
31.6	The Yukawa coupling for fermions: index structure	422
31.7	Smuggling mass into spin- $\frac{1}{2}$ fields: the Yukawa coupling	424
31.8	Smuggling mass into spin-1 fields: the scalar kinetic term	425
31.9	The full electroweak Higgs: a complex $SU(2)_L$ doublet	425
31.10	The reparametrisation: unitary gauge and Goldstone eating	427
31.11	The Fock vacuum and the mass gap	428

32 Fock Space for Gauge Fields	430
32.1 Enriching the single-particle state: direct products	430
32.2 The full single-particle state	431
32.3 The Standard Model Fock space	431
32.4 Creation and annihilation operators with full indices	432
32.5 The Fock hidden state and its evolution	433
33 Mode Expansion for Gauge Fields	434
33.1 Recap: the scalar mode expansion and its logic	434
33.2 Dirac field: spin and particle/antiparticle	434
33.3 Gauge boson fields: helicity and the removal of unphysical modes	435
33.4 Gluon field: colour-adjoint helicity modes	436
33.5 The W and Z bosons: massive vector fields	437
33.6 Summary: the mode expansion dictionary	438
33.7 The spin-statistics theorem: why the choice is forced	438
34 From Lagrangian Core to Hamiltonian Motor: the Schrödinger RNN	443
34.1 From Lagrangian to Hamiltonian: the Legendre transform	443
34.2 The Hamiltonian as a polynomial in creation and annihilation operators	443
34.3 The Fock vector as a snapshot of the universe	445
34.4 Energy, mass, and rotation of the Fock vector	446
34.5 The vacuum state	447
34.6 Fermion and boson fields	448
34.7 QED: the electromagnetic game engine	452
34.8 Electroweak theory: the symmetry-breaking game engine	455
34.9 QCD: the strong game engine	459
35 The Schrödinger Unrolled: Feynman Diagrams from the Dyson Series	466
35.1 Splitting the Hamiltonian: free and interacting parts	466
35.2 The interaction picture: separating free and interacting dynamics	466
35.3 The Dyson series: Taylor expanding the time evolution	467
35.4 Feynman diagrams as terms in the Dyson series	468
35.5 The Standard Model as a story generator	468
35.6 What the diagrams are: Fock-layer bookkeeping	469
35.7 The complete rendering pipeline	470
35.8 The monosemantic Hamiltonian versus the polysemantic residual stream	471
35.9 Technical difficulties: what the Dyson series cannot yet do	472
36 Path Integral for the Standard Model	474
36.1 Recap: the scalar path integral as a matrix product	474
36.2 Fermionic Fock space and its eigenstates	474
36.3 Grassmann numbers: algebra and calculus	475
36.4 The Dirac path integral	478
36.5 The round trip: Lagrangian $\rightarrow$ Hamiltonian $\rightarrow$ time slicing $\rightarrow$ Lagrangian	479
36.6 The QED path integral	482
36.7 The electroweak path integral	484
36.8 The full Standard Model path integral	485
37 Gauge Orbits, Faddeev–Popov, and BRST	489
37.1 Overview: dimension reduction and augmentation	489
37.2 The overcounting problem	490
37.3 The Faddeev–Popov procedure	491

37.4	Gauge fixing conventions	493
37.5	BRST symmetry	493
37.6	Physical states as BRST cohomology	495
37.7	Slavnov–Taylor identities	497
37.8	Ghost loops and asymptotic freedom	497
37.9	Anomalies: when BRST fails	498
37.10	The Fock picture: BRST and the physical Hilbert space	500
37.11	Summary: from nonlinear to linear dimension reduction	501
38	Feynman Rules for the Standard Model	504
38.1	Propagators from the quadratic part	504
38.2	Vertices from the interaction Lagrangian	505
38.3	Wick contractions with index tracking	505
38.4	Colour factors: a worked example	506
38.5	Ghost loops and gauge-invariance preservation	506
38.6	The complete Standard Model Feynman rules	507
38.7	The Fock-Born diagnostic for Feynman diagrams	507
38.8	Worked example: $e^+e^- \rightarrow \mu^+\mu^-$ from diagram to cross section	508
38.9	Beyond tree level: what loop corrections change	510
38.10	Electroweak corrections and the Z resonance	511
39	Renormalization of the Standard Model: Path Integral Side	512
39.1	The generating functional on a lattice	513
39.2	BRST constrains the counterterms to 19	513
39.3	The one-loop effective action: sum over species	514
39.4	The rendering-stability workflow for the Standard Model	514
39.5	Conflations specific to the gauge theory	515
39.6	$\mathbf{A}_0$: vacuum energy and the cosmological constant	516
39.7	$\mathbf{A}_1$: mass renormalization and the hierarchy problem	516
39.8	$\mathbf{A}_2$: the beta functions of the Standard Model	517
39.9	Screening versus antiscreening	518
39.10	The running of α_s : a Born-layer observable	519
39.11	Gauge coupling unification	520
39.12	Born-layer tests of the Standard Model	520
39.13	The 19 parameters of the Standard Model	521
39.14	Dimensional regularization as a computational tool	522
39.15	The Fock-Born summary	522
40	Renormalization of the Standard Model: the Hamiltonian Side	523
40.1	Gauge invariance without ghosts: the Gauss law	523
40.2	Where the ghost contribution goes	524
40.3	The physical Standard Model Fock space	524
40.4	The Standard Model Hamiltonian	525
40.5	Feshbach–Löwdin and the SM beta functions	525
40.6	The rendering-stability workflow: Hamiltonian side	526
40.7	Hamiltonian versus path integral: the full comparison	527
40.8	The Standard Model in the Fock-Born framework: a summary	528
41	Cataloging the Standard Model	529
41.1	A field guide to the particles: terminology and classification	529
41.2	Hadrons as patterns in the QCD Fock space: a formal treatment	535
41.3	Representation notation: bold numbers, $\otimes$, and $\oplus$	540

41.4	The Standard Model miracles	543
41.5	Confinement	547
41.6	The colour flux tube	548
41.7	Glueballs	548
41.8	The effective string theory of the flux tube	549
41.9	Hadron scattering and Regge trajectories	551
41.10	Lattice QCD: the engine made explicit	553
41.11	The hierarchy problem: what it is and what it is not	558
41.12	The mass gap problem: what should be proved	563
42	General Relativity from the Action Principle	567
42.1	Massive particles on flat spacetime	567
42.2	Einstein’s happiest thought: let $g_{\mu\nu}$ vary	567
42.3	Curvature as regularization: the Einstein–Hilbert action	568
42.4	The combined saddle point: the Einstein equation	568
42.5	Einstein’s uneasiness dissolved	569
42.6	Wheeler’s dictum, corrected	569
42.7	Black holes: the Schwarzschild geometry completed	570
42.8	Dark energy and the accelerating expansion	571
42.9	The rendering postulate and the Fock-Born reading	574
42.10	The Schwarzschild solution: a Born-layer output	574
42.11	Gravitational waves: ripples in the Born-layer output	575
42.12	Coupling gravity to the Standard Model	577
42.13	General relativity as a rendered symmetry	577
43	Quantum Gravity	579
43.1	The semiclassical framework	579
43.2	Quantising the scalar field on $g_{\mu\nu}$	579
43.3	The Unruh effect	580
43.4	The Hawking effect	580
43.5	Black hole evaporation	582
43.6	The information paradox	583
43.7	The Bekenstein–Hawking entropy and the heat kernel	584
43.8	The semiclassical synthesis	584
43.9	The graviton: a spin-2 field on $\eta_{\mu\nu}$	585
43.10	Effective field theory of gravity	585
43.11	The problem of time	585
43.12	Existing approaches, examined from Fock-Born	586
43.13	Induced gravity: the Einstein–Hilbert action from the heat kernel	587
43.14	Bare-parameter sensitivity and asymptotic safety	590
43.15	The one-way valve	593
44	Emergence of Collective Modes	595
44.1	Composite observables of fundamental fields	595
44.2	The quasiparticle: a composite that propagates	596
44.3	Worked example: the plasmon from the Lindhard function	597
44.4	Worked example: the Cooper pair and the BCS condensate	598
44.5	The effective propagator: direct computation	600
44.6	Certification: the quantitative criteria	601
44.7	Entropy and the Born-layer measure of certification	602
44.8	The vacuum state and condensation	603
44.9	The HS transformation for scalars: Euclidean path integral	604

44.10	The Gaussian noise of the auxiliary field	605
44.11	Block diagonalisation for scalars: Hamiltonian	605
44.12	The correspondence: Euclidean noise and Hamiltonian vacuum	606
44.13	The HS transformation for fermions	607
44.14	Block diagonalisation for fermions: Hamiltonian	608
44.15	The two cases compared	608
44.16	The fate of the auxiliary field	609
44.17	The origin of temperature: partial access	609
44.18	The Euclidean periodic time	610
44.19	The HS transformation at finite temperature	611
44.20	The gap equation and equation of state	611
44.21	Thermodynamics versus hydrodynamics: the Euclidean/Minkowski choice	612
45	String Theory and Loop Quantum Gravity	614
45.1	From hadrons to gravitons: the historical arc	614
45.2	The worldsheet action	614
45.3	Conformal field theory on the worldsheet	615
45.4	Vertex operators	616
45.5	The critical dimension	617
45.6	The string spectrum and the graviton	619
45.7	Scattering amplitudes: from worldsheet to spacetime	621
45.8	UV finiteness: modular invariance	622
45.9	Einstein–Hilbert from worldsheet consistency	624
45.10	The Kalb–Ramond field	625
45.11	Perturbative definitions and the search for M-theory	625
45.12	D-branes	628
45.13	AdS/CFT: the holographic correspondence	634
45.14	The Ryu–Takayanagi formula	637
45.15	Tensor networks and the geometry of entanglement	640
45.16	First and second quantisation of strings	643
45.17	Compactification: from ten dimensions to four	647
45.18	Moonshine: finite groups and modular forms	652
45.19	Why go beyond the metric?	656
45.20	The spin connection	657
45.21	The curvature of the spin connection	657
45.22	The Palatini action	658
45.23	The Holst action and the Barbero–Immirzi parameter	658
45.24	The 3 + 1 decomposition and Ashtekar variables	658
45.25	Quantisation: holonomies and fluxes	659
45.26	Spin networks: the kinematic Hilbert space	660
45.27	Area and volume quantisation	660
45.28	Fixing the Barbero–Immirzi parameter	661
45.29	Dynamics: constraints and spin foams	661
45.30	The diagnostic question	663
45.31	String theory: the graviton as a Born-layer object	663
45.32	Loop quantum gravity: geometry as a Born-layer object	664
45.33	The common diagnosis	665
45.34	The resolution: the engine is not gravitational	666
46	The Lattice-First View from Computational Science	668
46.1	Examining QFT problems on the lattice	668
46.2	Quantum foundation and HEP viewed from computational sciences	674

46.3	The two Clay problems: a lattice-first reading	678
46.4	The source code and the KISS principle	683
46.5	Numerical relativity is more than an approximation	688
47	Appendix: But What Is the Observer?	692
47.1	Schrödinger’s journey: from realism to Vedanta	692
47.2	Bohr’s discipline: the question that must not be answered	696
47.3	Heisenberg’s Platonism: potentiality and mathematical form	696
47.4	Three positions, one silence	697

1 Preface: The Fock-Born Framework and the Lattice-First View

This book is built on two layers, and two layers only:

Fock	The hidden state h_t : a unit vector in a Fock space $\mathcal{F}$, evolving under the Schrödinger equation $\dot{h}_t = -iHh_t$. This is the game engine: the complete, exact, deterministic description of the quantum world. It is the <i>potentiality</i> —everything the system could be, encoded as a superposition of basis vectors with complex amplitudes.
Born	The rendering layer: an observer submits an observable $\hat{O}$ to the Fock hidden state and receives a classical outcome—an eigenvalue of $\hat{O}$ —with a probability determined by the hidden state. This is the screen: the <i>actuality</i> that the observer records. It is irreducibly probabilistic, irreducibly classical, and irreducibly separate from the Schrödinger layer.

This is the structure of a recurrent neural network (RNN). The Fock layer is the hidden layer and the Born layer is the output layer.

More precisely, the Fock side is the *Fock-Schrödinger* layer (the hidden state and its unitary evolution), and the Born side is the *Born-Bohr* layer (the probabilistic rendering and the post-measurement update). We call the pair *Fock-Born* for short.

These two layers are sufficient to encompass the whole of quantum mechanics, condensed matter physics, and particle physics. Nothing else is needed at the foundational level. Everything else—Lagrangians, path integrals, Feynman diagrams, gauge fixing, BRST, renormalization—is a calculational tool for computing properties of h_t and matrix elements of e^{-iHT} , useful and powerful but not foundational.

The Fock side: a hierarchy of richer and richer basis vectors

The Fock space $\mathcal{F}$ is not a single fixed object. It is a hierarchy of increasingly rich Hilbert spaces, each obtained from the previous by adding indices to the basis vectors:

Level 1: a single spinless particle. Basis vectors $\{|\mathbf{x}\rangle\}$ or $\{|\mathbf{k}\rangle\}$, related by the Fourier transform. The hidden state is a wavefunction $\psi(\mathbf{x}, t) = \langle \mathbf{x} | h_t \rangle$. The Hamiltonian is $H = \mathbf{p}^2/2m + V(\mathbf{x})$. This is single-particle non-relativistic quantum mechanics.

Level 2: a single particle with spin. Basis vectors $\{|\mathbf{x}, s\rangle\}$ with spinor index $s \in \{1, 2\}$ (Pauli) or $A \in \{1, 2, 3, 4\}$ (Dirac). The Hamiltonian acts on the spinor index via Pauli or gamma matrices. Adding a spinor index is not a new formalism; it is the same Fock structure with one more label on the basis vector.

Level 3: n identical particles with fixed n . Basis vectors $\{|\mathbf{x}_1, s_1; \dots; \mathbf{x}_n, s_n\rangle_{\pm}\}$, symmetrised for bosons (+) and antisymmetrised for fermions (−). The Pauli exclusion principle is a theorem of the antisymmetry, not a postulate. Particle number is fixed; the Schrödinger equation preserves it.

Level 4: varying particle number (non-relativistic QFT). Basis vectors $\{|n; \mathbf{k}_1, s_1; \dots; \mathbf{k}_n, s_n\rangle\}$ for all $n = 0, 1, 2, \dots$, organised into the Fock direct sum $\mathcal{F} = \bigoplus_n \mathcal{H}^{(n)}$. Creation and annihilation operators connect adjacent sectors. The Hamiltonian is a polynomial of these operators. Phonons, magnons, Cooper pairs, and Bose–Einstein condensates all live here. This is non-relativistic quantum field theory—the natural language of condensed matter physics. The path integral and Lagrangian are optional: in condensed matter, one often writes the Fock Hamiltonian directly and never passes through a Lagrangian.

Level 5: relativistic fields. The basis vectors acquire Lorentz spinor indices, forced by the requirement that the Lagrangian be a Lorentz scalar and the field transform equivariantly under the Lorentz group. The Lagrangian is introduced here—not as a foundation, but as a device for writing Lorentz-invariant Hamiltonians efficiently. The path integral is a calculational tool for computing matrix elements of e^{-iHT} via time slicing. For fermions, the path integral variables are Grassmann numbers: anticommuting formal variables encoding the fermionic statistics.

Level 6: the Standard Model. The basis vectors carry the full set of indices: momentum $\mathbf{k}$, spin s , colour $\alpha \in \{r, g, b\}$, isospin $i \in \{1, 2\}$, species f , and particle/antiparticle label. The Hamiltonian H_{SM} is a polynomial of creation and annihilation operators, constrained by $SU(3)_c \times SU(2)_L \times U(1)_Y$ gauge invariance and Lorentz invariance, with 19 free parameters.

At each level, the Fock space is richer (more indices on the basis vectors), and the Hamiltonian is more constrained (more symmetries to satisfy). But the architecture is the same at every level: a unit vector in a Hilbert space, evolving unitarily under $\dot{h}_t = -iHh_t$. The Fock side is the game engine.

Now we turn to the screen.

The Born side: degeneracy is the key

The Born rendering layer converts the Fock hidden state into classical outcomes. The key to the entire Born side is a single concept: *degeneracy*.

An observable $\hat{O}$ is a Hermitian operator with eigenvalues $\lambda_1, \lambda_2, \dots$. Each eigenvalue may have *multiplicity*: many linearly independent eigenvectors may share the same eigenvalue. The set of all eigenvectors with eigenvalue λ spans an *eigensubspace* $\mathcal{H}_\lambda$. The *degeneracy* of λ is the dimension $d_\lambda = \dim \mathcal{H}_\lambda$: how many independent states the observable cannot distinguish.

The Born rule, stated in full generality, has three steps:

1. **Project.** Project h_t onto the eigensubspace $\mathcal{H}_\lambda$:

$$h_t^{(\lambda)} = P_\lambda h_t = \sum_{k=1}^{d_\lambda} |\lambda, k\rangle \langle \lambda, k | h_t \rangle, \quad (1.1)$$

where $\{|\lambda, k\rangle\}_{k=1}^{d_\lambda}$ is an orthonormal basis of $\mathcal{H}_\lambda$. This is the *projected superposition*: the piece of h_t compatible with eigenvalue λ . If $d_\lambda > 1$, it is a superposition of multiple eigenvectors—not a single state.

2. **Compute the probability.** The Born probability is the squared length of the projected superposition:

$$P(\lambda) = \|h_t^{(\lambda)}\|^2 = \sum_{k=1}^{d_\lambda} |\langle \lambda, k | h_t \rangle|^2. \quad (1.2)$$

The sum over k is the *OR rule*: the probability of outcome λ is the sum of the individual probabilities for each eigenvector in the eigensubspace. Cross terms vanish by orthogonality—no interference within the eigensubspace.

3. **Trim and renormalize (Bohr collapse).** If outcome λ is observed, the hidden state is updated to the projected superposition, renormalized to unit length:

$$h_{t+} = \frac{h_t^{(\lambda)}}{\|h_t^{(\lambda)}\|} = \frac{P_\lambda h_t}{\sqrt{P(\lambda)}}. \quad (1.3)$$

Everything outside $\mathcal{H}_\lambda$ is trimmed away. Everything inside survives with its original amplitudes and phases. The Schrödinger evolution resumes from h_{t+} .

The degeneracy d_λ controls everything about the measurement:

- If $d_\lambda = 1$ (non-degenerate): the projected superposition is a single eigenvector. The collapse sets $h_{t+} = |\lambda\rangle$ —complete trimming. This is the textbook Born rule.
- If d_λ is large: the projected superposition retains a rich superposition inside $\mathcal{H}_\lambda$. The measurement trims little and learns little.
- If $d_\lambda = \dim \mathcal{H}$ (the eigenvalue fills the entire space): the projected superposition is h_t itself. Nothing is trimmed. Nothing is learned.

The identity observable: the most revealing example. The simplest observable is the identity operator $\hat{O} = \mathbf{1}$. It has a single eigenvalue $\lambda = 1$ with degeneracy $d_1 = \dim \mathcal{H}$: every vector is an eigenvector. The projected superposition is h_t itself. The probability is $P(1) = \|h_t\|^2 = 1$. The collapse does nothing: $h_{t+} = h_t$.

The identity asks no question, distinguishes no states, and learns nothing. It is the maximally degenerate extreme—the measurement that leaves the hidden state unchanged. Every other observable is a partial lifting of this degeneracy: it splits the eigenspace of the identity into subspaces, distinguishes some states, and trims some of the superposition. The identity is the starting point; all observables are refinements of it.

The Born hierarchy: levels of degeneracy

The same Fock hidden state h_t can be rendered by different observables. The *degeneracy* of the observable determines the character of the output:

Observable	Degeneracy	Projected position	super- ance	Outcome	vari-	Character
Identity $\mathbf{1}$	$d = \dim \mathcal{H}$	h_t itself		Zero		No measurement
Macroscopic	$d \sim e^S$	Huge subspace		$\sim 1/\sqrt{N} \rightarrow 0$		Classical
Few-body	$d \sim O(1)$	Small subspace		$O(1)$		Quantum
Complete set	$d = 1$	Single vector		Maximal		Fully quantum

We now describe each level, from the finest measurement to the coarsest.

Level 1: non-degenerate observable ($d_\lambda = 1$). Every eigenvalue has a unique eigenvector. The projected superposition is a single basis vector. The Born probability is $P(\lambda) = |\langle \lambda | h_t \rangle|^2$ —the bare Born rule with no OR sum. The Bohr collapse sets $h_{t+} = |\lambda\rangle$ —complete trimming to a single state. Full quantum interference is visible. This is the regime of quantum information: trapped ions, superconducting qubits, single photons.

Level 2: partial measurement (moderate degeneracy). The system S is entangled with an environment E ; the observer measures $\hat{O}_S \otimes \mathbf{1}_E$ (an observable on S tensored with the identity on E). The identity $\mathbf{1}_E$ has degeneracy $\dim \mathcal{H}_E$ —the entire environment is unresolved. The projected superposition onto eigenvalue λ_S sums over all environment states:

$$h_t^{(\lambda_S)} = \sum_e \langle \lambda_S, e | h_t \rangle |\lambda_S, e\rangle. \quad (1.4)$$

The Born probability $P(\lambda_S) = \sum_e |c(\lambda_S, e)|^2$ is the OR sum over the environment—the interference between different environment states is destroyed by the degeneracy. This is decoherence:

the environment's degeneracy washes out quantum interference, making the Born output appear classical.

Temperature emerges here. The reduced density matrix of a system entangled with a large generic environment is the Gibbs state $\rho_S = e^{-\beta H_S}/Z$, by the *eigenstate thermalisation hypothesis* (ETH). The ETH states that in a generic (non-integrable) many-body system, the expectation value of any local observable $\hat{O}_S$ in an energy eigenstate $|E_n\rangle$ of the full system $S + E$ equals the microcanonical average at that energy:

$$\langle E_n | \hat{O}_S | E_n \rangle \approx \bar{O}_S(E_n) \equiv \frac{1}{\mathcal{N}(E_n, \Delta E)} \sum_{\substack{m: \\ |E_m - E_n| < \Delta E}} \langle E_m | \hat{O}_S | E_m \rangle, \quad (1.5)$$

where $\mathcal{N}(E_n, \Delta E)$ is the number of eigenstates in the energy window. The off-diagonal matrix elements $\langle E_m | \hat{O}_S | E_n \rangle$ ($m \neq n$) are exponentially small in the system size: they scale as $e^{-S(E)/2}$, where $S(E) = \ln \mathcal{N}(E, \Delta E)$ is the microcanonical entropy.

The physical content: a single energy eigenstate of a large system looks thermal to a local observer. The observer who measures only $\hat{O}_S$ cannot distinguish the pure state $|E_n\rangle \langle E_n|$ from the microcanonical ensemble—the exponentially large degeneracy of the environment hides the eigenstate identity behind the OR sum. Temperature is the Born rendering of the system observable with the environment traced out—a measure of the degeneracy of the unobserved environment.

In ML terms: ETH is the statement that a single hidden state of the game engine, when read out through a low-resolution observable (one that sees only a small subsystem), produces outputs indistinguishable from thermal noise. The “randomness” is not in the engine—the engine is in a definite eigenstate, rotating deterministically—but in the coarse-graining of the readout.

Level 3: statistical mechanics (high degeneracy). $P(\lambda) = e^{-\beta E_\lambda}/Z$: the thermal Born rule. This is Level 2 in the thermodynamic limit, when the environment is large and generic. Statistical mechanics is the Born rule applied to a system whose environment has degeneracy $\sim e^{S_{\text{env}}}$ —an enormous OR sum that produces the Boltzmann distribution.

Level 4: macroscopic observables (extreme degeneracy). The observer measures a collective observable $\hat{\Phi} = N^{-1} \sum_{k=1}^N \hat{\phi}_k$ over $N \sim 10^{23}$ degrees of freedom. Each eigenvalue of $\hat{\Phi}$ has degeneracy $\sim e^S$, where S is the thermodynamic entropy.

The Born variance is suppressed by $1/N$. To see why, compute the variance of $\hat{\Phi}$ from the Born rule. Each $\hat{\phi}_k$ has expectation $\langle \hat{\phi}_k \rangle = \mu$ and variance $\text{Var}(\hat{\phi}_k) = \sigma^2$ (both Born-layer quantities, rendered from h_t). Because the OR sum over the unresolved environment destroys cross-correlations between different sites (decoherence), the Born covariance $\text{Cov}(\hat{\phi}_j, \hat{\phi}_k) \rightarrow 0$ for $j \neq k$ in the decoherent regime. The variance of the average is then:

$$\text{Var}(\hat{\Phi}) = \text{Var}\left(\frac{1}{N} \sum_{k=1}^N \hat{\phi}_k\right) = \frac{1}{N^2} \sum_{k=1}^N \text{Var}(\hat{\phi}_k) = \frac{\sigma^2}{N} \xrightarrow{N \rightarrow \infty} 0. \quad (1.6)$$

This is the law of large numbers applied to Born outputs. For $N \sim 10^{23}$: $\sigma/\sqrt{N} \sim 10^{-12}\sigma$. The Born distribution of $\hat{\Phi}$ is a spike of width $\sim \sigma/\sqrt{N}$, indistinguishable from a delta function on any experimental scale. Sampling from a delta distribution is deterministic. Trimming a delta distribution leaves h_t unchanged. The Born layer is inert: the classical value is the only outcome.

The condensate order parameter $\langle \hat{\psi}(\mathbf{x}) \rangle = \sqrt{N_0/V} e^{i\theta}$ is the cleanest example: $N_0 \sim N$ particles in the same Fock mode. The number operator $\hat{n} = a^\dagger a$ in this mode has Born mean $\bar{n} = N_0$ and Born variance $\text{Var}(\hat{n}) \sim N_0$ (Poissonian for a coherent state), so the relative fluctuation is

$\sqrt{\text{Var}(\hat{n})}/\bar{n} = 1/\sqrt{N_0} \rightarrow 0$. The field value $\hat{\psi}(\mathbf{x})$ is rendered as a deterministic classical field—a Born output with negligible variance.

Classical mechanics, thermodynamics, hydrodynamics, and condensate physics all live at Level 4: the regime where the degeneracy is so extreme that the Born output has negligible variance.

The consistency chain

The four Born levels are not four different theories. They are one theory—the Born rule applied to one Fock hidden state—at four different degeneracies. The consistency chain:

$$\underbrace{|\langle \lambda | h_t \rangle|^2}_{\substack{\text{Level 1} \\ d_\lambda=1}} \xrightarrow{\text{trace out } E} \underbrace{\sum_e |c(\lambda, e)|^2}_{\substack{\text{Level 2} \\ d \sim \dim \mathcal{H}_E}} \xrightarrow{\text{ETH}} \underbrace{e^{-\beta E_\lambda}/Z}_{\substack{\text{Level 3} \\ d \sim e^{S_{\text{env}}}}} \xrightarrow{N \rightarrow \infty} \underbrace{\langle \hat{\Phi} \rangle \text{ determ.}}_{\substack{\text{Level 4} \\ d \sim e^{S_{\text{thermo}}}}} \quad (1.7)$$

Each arrow is a well-defined mathematical operation, and no new physics is introduced at any step:

The first arrow (*trace out E*) is the projected superposition: the Born rule with the environment’s identity as the degenerate factor. The OR sum over environment states $\sum_e |c(\lambda_S, e)|^2$ is a marginalisation—the quantum analogue of summing a joint probability over one variable.

The second arrow (*ETH*) replaces the OR sum with the Boltzmann weight. The ETH, stated in (1.5), guarantees that the diagonal matrix elements of $\hat{O}_S$ in energy eigenstates are smooth functions of energy, so the OR sum $\sum_e |c(\lambda, e)|^2$ over the $e^{S_{\text{env}}}$ environment states is well-approximated by the microcanonical average, which for a large environment equals the canonical (Gibbs) distribution $e^{-\beta E_\lambda}/Z$ with $\beta = \partial S_{\text{env}}/\partial E$. This is the standard derivation of the canonical ensemble from the microcanonical one, here expressed as a property of the Born-layer OR sum.

The third arrow (*law of large numbers*) is the variance suppression (1.6): when $N \rightarrow \infty$ degrees of freedom are averaged, the Born distribution concentrates into a delta spike, and the output becomes deterministic.

The degeneracy increases at each level; the variance decreases; and the Born output becomes progressively more classical. The chain is not a sequence of approximations applied to different theories. It is a single Born rule, applied to a single Fock hidden state, rendered through observables of increasing degeneracy. The classical world is the rendering at Level 4—the regime where the Born distribution has already decided.

There is no quantum-to-classical transition

The Fock hidden state is always quantum: a complex unit vector in Hilbert space, evolving unitarily under $\dot{h}_t = -iHh_t$. The Born output is always classical: a real eigenvalue—a definite datum recorded by the observer. There is no moment at which one converts into the other. The architecture has both layers from the start: the Fock layer is the game engine, the Born layer is the screen, and neither ever becomes the other.

What varies from observable to observable is the Born probability distribution over the eigenvalues—its concentration, its entropy, its randomness.

For a macroscopic observable $\hat{\Phi} = N^{-1} \sum_k \hat{\phi}_k$ acting on $N \sim 10^{23}$ degrees of freedom, the degeneracy of each eigenvalue is enormous ($d_\lambda \sim e^S$), and the OR sum over each eigensubspace concentrates the Born distribution into a near-delta spike: variance $O(1/N)$, entropy negligible. Born sampling from a near-delta distribution is trivial—every sample returns essentially the same value. The Bohr collapse barely alters the hidden state, because the projected superposition

is nearly h_t itself. This is the classical regime: not because the Fock layer has stopped being quantum, but because the Born distribution is so concentrated that its randomness is invisible and its collapse is inert. The observer need not sample; the distribution has already decided.

For a non-degenerate observable ($d_\lambda = 1$ for every λ), the Born distribution can spread across many eigenvalues with comparable amplitudes—high entropy, genuine randomness. Each Born sample is unpredictable. Interference is fully visible. The Bohr collapse reshapes the hidden state drastically, trimming it to a single eigenvector. This is the quantum regime.

The identity $\hat{O} = \mathbf{1}$ sits at one pole: one eigenvalue, probability one, zero entropy, zero information, zero collapse. A complete set of commuting observables sits at the other: every eigenvalue non-degenerate, maximal resolution, full collapse. Every physical measurement lies between these two poles.

There is no quantum-to-classical transition. There is a single architecture—Fock hidden state, Born rendering—and a continuous gradient in the entropy of the Born distribution, controlled by the degeneracy of the observable the observer chooses to submit to the game engine.

The path integral is a calculation tool, not a foundation

The path integral lives entirely inside the Fock layer. It is a method for computing the matrix element $\langle \phi_f | e^{-iHT} | \phi_i \rangle$ by time-slicing: factor e^{-iHT} into N small steps, insert a resolution of the identity $\mathbf{1} = \int d\phi |\phi\rangle \langle \phi|$ at each step, and obtain

$$\langle \phi_f | e^{-iHT} | \phi_i \rangle = \int \prod_{k=1}^{N-1} d\phi_k \prod_{k=0}^{N-1} \langle \phi_{k+1} | e^{-iH\epsilon} | \phi_k \rangle \xrightarrow{N \rightarrow \infty} \int \mathcal{D}\phi e^{iS[\phi]}. \quad (1.8)$$

The field value $\phi(\mathbf{x}, t_k)$ at each slice is an eigenvalue of $\hat{\phi}(\mathbf{x})$ —but it is a *bookkeeping eigenvalue*, a basis label inserted to decompose a Fock-layer matrix element. It is not the outcome of a Born-layer measurement. No observer samples it. No Bohr collapse occurs at t_k . The system does not “visit” the configuration $\phi(\mathbf{x}, t_k)$. The Lagrangian L re-emerges from the time-sliced Hamiltonian by Gaussian integration over the conjugate momentum—a mathematical step inside the Fock layer, not a physical principle and not a rendering.

The self-tripping temptation. It is extremely tempting to reinterpret these bookkeeping eigenvalues as Born-layer eigenvalues. The field operator $\hat{\phi}(\mathbf{x})$ is, after all, a legitimate Hermitian operator. For a macroscopic field it is a highly degenerate observable whose Born distribution is near-delta: if you *measured* $\hat{\phi}(\mathbf{x})$, you would get a sharp value with near-certainty. So the reasoning goes: “the field really does have the value $\phi(\mathbf{x}, t)$ at each time; the path integral is a sum over the actual histories the system might take.”

This slides from the Fock layer into Born-layer language without noticing the crossing. The bookkeeping eigenvalue at time slice t_k is a dummy index in a matrix trace. The Born eigenvalue of a macroscopic observable is the outcome of a measurement act—a rendering through the Born layer, with projection, sampling, and collapse. These are different operations in different layers of the architecture, and conflating them is the central confusion of the “sum over histories” picture.

Feynman’s “the electron takes every path” and Hawking’s “sum over compact geometries” both invite exactly this conflation. The word *history* implies a sequence of actualities—Born-layer events—but each “history” is a Fock-layer basis label, no more actual than a row index in a matrix multiplication. The picture is vivid and computationally powerful, but it is self-tripping: the formalism’s own language seduces the user into promoting a calculational tool to an ontological foundation.

The Grassmann reminder. For fermions, the integration variables are Grassmann numbers: formal anticommuting symbols with no classical meaning. There is no fermionic field configuration that the system “visits”—the notion is algebraically impossible. This should serve as a standing warning that path-integral variables are computational devices, not physical states—even in the bosonic case, where the temptation to reify them is stronger because the eigenvalues happen to be real numbers.

Condensed matter as the healthy model. In condensed matter physics, one writes the Fock Hamiltonian directly—Hubbard, Heisenberg, BCS, Bose–Hubbard, Kondo—in terms of creation and annihilation operators, and never passes through a Lagrangian. The path integral appears when needed (Hubbard–Stratonovich transformations, saddle-point approximations, large- N limits) and is recognised for what it is: a calculational detour from the Hamiltonian, not a replacement for it. The Hamiltonian is always there to return to.

Particle physics, by historical accident, learned quantum field theory through Feynman diagrams and the Lagrangian path integral, and this created a culture in which $\mathcal{L}$ appears foundational and H derivative. The Fock-Born framework reverses the priority: the Lagrangian is the scaffold; the Fock Hamiltonian is the building.

The correct physical intuition

The correct physical picture has two elements:

1. **The Fock layer.** A unit vector h_t in Fock space, rotating under the Hamiltonian: $h_t = e^{-iHt}h_0$. The rotation is continuous, deterministic, and unitary. At each instant the hidden state is a single vector—not a collection of histories, not a swarm of particles, not a fluctuating field. A single vector, rotating.
2. **The Born layer.** An observable $\hat{O}$, which provides two things: a *vocabulary* of eigenvalues $\{\lambda\}$ —the possible classical outcomes—and a *readout rule* via the eigenvectors $\{|\lambda, k\rangle\}$, which tells how to extract the probability of each outcome from the hidden state by projection and squared length.

There is nothing else.

No further mechanism, no additional layer, no hidden classical substratum is needed. Every additional picture that has entered the folklore of quantum physics is, in one way or another, a Born-layer output pushed backwards into the Fock layer—an effect misread as a cause, a rendering confused with the engine. The Fock-Born framework provides a systematic diagnostic for identifying and dissolving these conflations.

We now catalogue the most important ones.

“Sum over histories.” Covered in the preceding subsection: the path integral is a Fock-layer matrix computation—a method for evaluating $\langle\phi_f|e^{-iHT}|\phi_i\rangle$ —not a sum over Born-layer actualities. The field value at each time slice is a bookkeeping eigenvalue, not a measured outcome. The system does not take every path. The hidden state is one vector, not many histories.

“Virtual particles.” Every internal line in a Feynman diagram is called a “virtual particle.” The name itself is the error: *particle* is a Born-layer word—a detected eigenvalue of a number operator or a momentum observable—and nothing in the Fock-layer computation produces a detection event. The internal line is a contraction—a matrix element of the free propagator $\langle 0|T\{\hat{\phi}(x)\hat{\phi}(y)\}|0\rangle$ —appearing inside a perturbative expansion of $\langle f|e^{-iHT}|i\rangle$. It is a bookkeeping device, a Fock-layer index summed over, not a Born-layer outcome.

“Exchange of particles.” The Feynman diagram for electron–electron scattering shows a photon line connecting the two fermion lines. The pictorial language says: “electron 1 emits a photon; electron 2 absorbs it.” This narrates a sequence of Born-layer events—emission, propagation, absorption—as if they were observed actualities. In the Fock layer, what happens is simpler: the interaction Hamiltonian $H_{\text{int}} = e \int d^3x \bar{\psi} \gamma^\mu \psi A_\mu$ has a non-zero matrix element between the initial two-electron Fock state and the final two-electron Fock state. The “exchanged photon” is a term in the perturbative expansion of this matrix element—an intermediate Fock-space basis label summed over, not a particle that travels from one electron to the other. No observer detects it. No Born rendering occurs at the internal vertex.

“The vacuum teems with virtual particles.” The vacuum $|0\rangle$ is the ground state of the full Hamiltonian H . It is a single, definite Fock vector—the lowest-energy eigenstate, sitting still (up to a global phase $e^{-iE_0 t}$). It does not teem. The “teeming” language arises from expanding $|0\rangle$ in the *free-field* Fock basis:

$$|0\rangle_{\text{full}} = c_0 |0\rangle_{\text{free}} + \sum_{\mathbf{k}} c_{\mathbf{k}} |1_{\mathbf{k}}\rangle + \sum_{\mathbf{k}, \mathbf{k}'} c_{\mathbf{k}\mathbf{k}'} |1_{\mathbf{k}}, 1_{\mathbf{k}'}\rangle + \dots \quad (1.9)$$

Because the interacting vacuum differs from the free vacuum, the expansion has components in multi-particle sectors. These components are basis-decomposition coefficients—Fock-layer projections onto a particular basis—not Born-layer events. Nothing is being created or annihilated. The vacuum is one vector, and it is quiet.

“Tunnelling through the barrier.” The standard narrative: “the particle tunnels through the classically forbidden region.” This narrates a Fock-layer rotation as if it were a Born-layer trajectory. In the Fock layer: the Hamiltonian $H = p^2/2m + V(x)$ has eigenstates that are nonzero on both sides of the barrier. The symmetric and antisymmetric combinations $|+\rangle$ and $|-\rangle$ are split by an energy gap ΔE . The time evolution e^{-iHt} rotates the hidden state from one with high amplitude on the left to one with high amplitude on the right, with period $\pi\hbar/\Delta E$. No trajectory through the barrier occurs. The hidden-state vector simply rotates in Hilbert space: the amplitude shifts continuously from left-localised basis vectors to right-localised basis vectors, because the eigenstates of H span both sides.

The instanton—the classical path in imaginary time with action S_E , giving a splitting $\Delta E \propto e^{-S_E/\hbar}$ —is a path-integral computation of this energy gap. It is a useful Fock-layer calculation, not a Born-layer journey. The phrase “tunnels through” splices two Born-layer position measurements (position left at t_1 , position right at t_2) into a Fock-layer narrative of travel, as if the system visited the forbidden region in between. It did not. It rotated.

“Quantum fluctuations.” Perhaps the most pervasive conflation. “The field fluctuates.” “The vacuum fluctuates.” “Quantum fluctuations seed cosmic structure.” The word *fluctuation* implies a quantity changing in time—Born-layer language. In the Fock layer, the vacuum $|0\rangle$ is a single definite vector, sitting still. It does not fluctuate. What the word actually refers to is Born-layer *variance*: if you measure the field operator $\hat{\phi}(\mathbf{x})$ repeatedly on identically prepared vacuum states, the outcomes scatter— $\langle 0|\hat{\phi}^2|0\rangle \neq 0$ —but this is a statistical property of the Born distribution, not a temporal process in the Fock layer. The conflation treats Born-layer variance as Fock-layer motion.

In inflationary cosmology, “quantum fluctuations become classical perturbations” means: the Born distribution of each inflaton mode amplitude becomes highly squeezed (near-delta) after horizon exit, because the mode’s Fock-layer evolution under H drives the state into a highly squeezed state. The Fock state does not fluctuate and then freeze. The Born statistics change character as the mode evolves unitarily.

“The atom emits a photon” (spontaneous emission). The narrative implies an event at a definite time: the atom decides to emit, and a photon is born. In the Fock layer: the interaction Hamiltonian couples $|\text{excited}, 0_\gamma\rangle$ to $|\text{ground}, 1_{\mathbf{k}\lambda}\rangle$, and e^{-iHt} continuously rotates the hidden state from the former toward a superposition including the latter. No event occurs at any particular moment. The Fock vector rotates smoothly. “Emission” is a Born-layer detection—the photon counter clicks—projected backward onto the Fock layer as a moment of creation. The Weisskopf–Wigner exponential decay law is a Born-layer statistical prediction: if you measure the atom’s state at time t , the probability of finding it still excited is $e^{-\Gamma t}$. The Fock-layer evolution is continuous and deterministic throughout.

“The Casimir effect proves virtual particles are real.” This is the flagship argument for reifying virtual particles: the force between the plates is measurable, therefore the vacuum fluctuations that produce it must be real. In the Fock layer: the Casimir energy is the difference in ground-state eigenvalue of the electromagnetic Hamiltonian with boundary conditions (plates) versus without. It is a property of one vector—the ground state $|0\rangle_{\text{plates}}$ —and its energy eigenvalue $E_0(d)$ depends on the plate separation d . No particles are created or annihilated. The force $F = -dE_0/dD$ is a Fock-layer energy derivative, computable without mentioning virtual particles at all—Lifshitz computed it from the electromagnetic Green’s function without any mode sum or vacuum-fluctuation language. The “proof” that virtual particles are real is a proof that the ground-state energy depends on boundary conditions—which it does, and which requires no ontological commitment beyond one quiet vector.

“Wave-particle duality.” The oldest conflation. “The electron sometimes behaves as a wave, sometimes as a particle.” In Fock-Born terms: the hidden state is always a single vector in Fock space. It is neither a wave nor a particle. “Wave” and “particle” are descriptions of Born-layer output patterns under different observables. Measure position with a non-degenerate detector: you get a localised click (particle-like Born output). Measure the interference pattern—a highly degenerate observable that resolves only fringe positions—and you get an interference distribution (wave-like Born output). The duality is a property of the Born layer’s observable choice, not of the Fock-layer state. The electron does not switch between two natures. The observer switches between two observables, and the Born distributions look different.

The deepest layer violation: $g_{\mu\nu}$ as a recycled Born-layer output. The Fock-Born architecture has a one-way flow:

$$\text{Fock layer} \xrightarrow{\text{Born rendering}} \text{classical outcomes (eigenvalues)} \quad (1.10)$$

The Born layer reads from the Fock layer. Its outputs are classical data—numbers recorded by observers. Once a quantity has been rendered as a Born-layer output, it cannot be fed back as a Fock-layer input without violating the directionality of the architecture. It is a one-way valve.

The preceding conflations are *interpretive* layer violations: they narrate Fock-layer computations in Born-layer words—particles emitted, paths taken, fields fluctuating, vacua teeming. They are confusions you can correct by choosing better words. The metric $g_{\mu\nu}$ is a *structural* layer violation: the Fock layer’s definition literally depends on a Born-layer quantity.

Operationally, $g_{\mu\nu}$ is what clocks, rods, and light signals report—the eigenvalue of a macroscopic gravitational observable with a near-delta Born distribution. Yet $g_{\mu\nu}$ is used to build the Fock layer itself: it defines the Lorentzian manifold, the mode decomposition that determines what “particle” means, the causal structure that orders events, and the time parameter in e^{-iHt} . The output of the rendering engine is being used to build the game engine.

In non-gravitational physics, the violation is harmless: the background $g_{\mu\nu} = \eta_{\mu\nu}$ is flat Minkowski space, and its Born-layer origin can be forgotten. The Born distribution is not merely near-delta

but exactly delta— $\eta_{\mu\nu}$ never fluctuates, never varies, never needs to be measured.

In quantum gravity, the violation becomes the problem—and *every* existing approach commits it, not only the path integral.

The path-integral approaches attempt to promote $g_{\mu\nu}$ to a dynamical field and integrate over it: $\int \mathcal{D}g e^{iS_{\text{EH}}[g]}$. But the path integral’s very definition—time-slicing e^{-iHT} , inserting resolutions of the identity—presupposes a background causal structure and a time parameter, both of which come from the metric. One is integrating over the object that defines the integration. Hawking’s sum over compact geometries, the Euclidean gravitational path integral, the string-theory landscape of vacua—all inherit this circularity.

The Hamiltonian approaches—Wheeler–DeWitt, loop quantum gravity—do not escape it by adopting the operator formalism. Wheeler–DeWitt quantises the 3-metric h_{ij} , but h_{ij} is $g_{\mu\nu}$: a Born-layer observable promoted to a Fock-layer variable. The lapse function that would tick the Schrödinger evolution is itself a component of $g_{\mu\nu}$ —a Born-layer clock recycled as the engine’s time parameter. The “problem of time” ($\hat{H}\Psi = 0$, with no external time parameter) is not a bug in the Hamiltonian approach. It is the *symptom* of the layer violation: the clock that should tick the Fock-layer evolution is itself part of what is being quantised, and the formalism seizes up because a Born-layer output cannot simultaneously serve as a Fock-layer foundation. Loop quantum gravity’s spin networks discretise spatial geometry, but the dynamics still requires a clock—and every candidate clock is $g_{\mu\nu}$ -derived. The Hamiltonian approach makes the layer violation *visible*; it does not resolve it.

This is a failure with a deep historical root. Bohr dismantled classical realism for every other observable: position, momentum, spin, energy—all were relocated from the world-in-itself to the Born layer, existing as definite values only when an observer submits an observable and receives an eigenvalue. But Bohr never touched $g_{\mu\nu}$. Spacetime geometry was left as the one classically real backdrop—the stage on which both quantum system and measuring apparatus perform, never itself subjected to the Born-layer analysis. The Copenhagen interpretation says “the measuring apparatus is classical,” and the spacetime that the apparatus inhabits was treated as even more fundamental: not merely classical but *unquestionably* classical, the last piece of classical realism left standing after the quantum revolution dismantled everything else.

The difficulty of quantum gravity, from the Fock-Born viewpoint, is therefore not “how to quantise $g_{\mu\nu}$ ” and not “path integral versus Hamiltonian.” It is that no existing approach—path-integral or Hamiltonian, perturbative or non-perturbative—has yet built a Fock layer whose architecture does not depend on the output of its own Born layer. The problem is the layer violation itself, and it is shared by all.

The dissolution: separate the engine from the rendering. The layer violation disappears once you stop using $g_{\mu\nu}$ to build the Fock layer. The move is simple in principle: let the Fock-layer computational substrate be a flat spacetime lattice—a discrete set of sites with spacing a , connected by nearest-neighbour links, carrying the degrees of freedom of the fields. The Hamiltonian H is a local operator on this lattice. The Fock vector h_t is a unit vector in the lattice Hilbert space, evolving under $h_t = e^{-iHt}h_0$. The lattice is not built on a manifold defined by $g_{\mu\nu}$; the lattice *is* the computational substrate—the game engine—and it needs no metric to define it.

The metric $g_{\mu\nu}$ then emerges where it belongs: in the Born layer. At wavelengths $\lambda \gg a$, the Born-layer observables of the lattice theory—correlation functions, propagation amplitudes, causal structure—organise themselves as if they lived on a smooth manifold with metric $g_{\mu\nu}$. The metric is rendered, not foundational. Curvature, geodesics, horizons, and the Einstein equations are all Born-layer descriptions of patterns in the long-wavelength rendering of a flat-lattice computation. The Fock layer knows nothing of curvature. It has sites, links, and a Hamiltonian.

From this viewpoint, the “deep” problems of quantum gravity—the problem of time, the non-renormalisability of G_N , the black-hole information paradox as standardly framed—are all symptoms of a single error: treating a Born-layer output ($g_{\mu\nu}$) as a Fock-layer foundation and then struggling with the consequences of the recycling. Once the engine is a lattice and the metric is a rendering, the problems dissolve as conflation. There is no “problem of time” because the lattice has its own time parameter—the discrete ticking of $e^{-iH\epsilon}$ —which does not depend on $g_{\mu\nu}$. There is no “non-renormalisability” because the lattice spacing a is the bottom of the engine, and no physics exists below it.

The computational substrate is discrete: the lattice-first view. This book takes the position that the Fock layer is a computational substrate—a game engine—and that a computational substrate must be discrete. The continuum quantum field theory that appears in textbooks is a long-wavelength approximation to the discrete lattice, not the other way around. The lattice is not a “regularisation” of a more fundamental continuum theory; the continuum is a *rendering* of the lattice, valid when $\lambda \gg a$.

This reversal of priority dissolves the conceptual tangle of renormalisation. The standard presentation runs: start with a continuum Lagrangian, discover that it produces divergences, impose a cutoff to regulate them, add counterterms to absorb them, and then take the cutoff to infinity to recover the “true” continuum theory. At every step, the continuum is treated as fundamental and the cutoff as an artefact to be removed.

The lattice-first viewpoint runs differently:

1. The Fock-layer Hamiltonian $H(a)$ is defined on a lattice with spacing a . It has bare parameters $\theta(a)$ —masses, couplings, field-strength normalisations—that are functions of a .
2. The Born layer renders observables: cross sections, masses, binding energies, correlation lengths. These are measured at wavelengths $\lambda \gg a$.
3. **Renormalisability** is the requirement that Born-layer observables are stable under changes of a , up to corrections that vanish as $a/\lambda \rightarrow 0$:

$$\mathcal{O}_{\text{Born}}(\lambda) = \mathcal{O}_{\text{phys}} + o\left(\frac{a}{\lambda}\right), \quad (1.11)$$

provided the bare parameters $\theta(a)$ are adjusted as a varies. The adjustment of $\theta(a)$ is what textbooks call “renormalisation.” In the lattice-first viewpoint, it is simply the statement that different lattice spacings can render the same long-wavelength physics—the engine can be re-meshed without changing the picture on the screen.

4. The “divergences” of the continuum theory are the statement that the bare parameters $\theta(a)$ grow without bound as $a \rightarrow 0$. In the lattice-first viewpoint, this is not a crisis; it is the unsurprising fact that a lattice parameter depends on the lattice spacing. The limit $a \rightarrow 0$ is a mathematical extrapolation, not a physical requirement. The lattice at finite a is the theory.

Renormalisability, in this light, is *evidence* that the computational substrate of the Fock layer is a lattice. It says: Born-layer observables are insensitive to the engine’s mesh size, which is exactly what you expect if the engine is discrete and the rendering operates at wavelengths far above the mesh. A renormalisable theory is one whose rendering is stable. A non-renormalisable theory is one whose rendering depends on the mesh in a way that cannot be absorbed into a finite number of bare parameters—it needs the full specification of the lattice Hamiltonian, which is not a sickness but simply a theory with more parameters.

The lattice dissolves the rigour problem. The programme of constructive quantum field theory—Glimm, Jaffe, and their successors—sought to rigorously define the continuum path

integral measure $\mathcal{D}\phi$ and to construct interacting quantum field theories as mathematical objects. This is a programme about the rendering, not the engine. On a finite lattice with spacing a , the path integral is a finite-dimensional ordinary integral—no measure-theoretic subtlety arises. The Hamiltonian is a well-defined operator on a separable Hilbert space—no domain issues, no distributional ambiguities. Both the path integral and the Hamiltonian are rigorously defined at finite a , and no continuum limit is needed for the physics.

The rigour problem is an artefact of demanding that the rendering be promoted to a self-standing mathematical object, independent of the engine that produces it. If the engine is a lattice, the rendering is always derived, and the demand is misplaced.

There is no UV completion. If the lattice spacing a is the bottom of the Fock-layer engine, there is nothing below it. The question “what is the UV completion of this theory?” asks about the structure of the engine below its smallest component. But the engine has no structure below a . The sites and links at scale a are the finest grain of the computational substrate. UV completion is not a well-defined observable; it is a Born-layer extrapolation pushed below the engine’s resolution, and there is nothing there to find.

Asymptotic safety—Weinberg’s proposal (1979) that gravity or other non-renormalisable theories might possess a non-trivial UV fixed point, making the continuum limit $a \rightarrow 0$ well-defined—is a question about the rendering, not the physics. It asks: can the bare parameters $\theta(a)$ be tuned so that they flow to a fixed point as $a \rightarrow 0$, producing a self-consistent continuum theory? From the lattice-first viewpoint, this is asking whether the screen image can be sharpened to infinite resolution. The answer may be mathematically interesting, but it is physically irrelevant: the engine runs at finite a , and no observable probes the fixed point. (Asymptotic *freedom*, by contrast, is physically real: the running of $\alpha_s(\mu)$ at different energy scales μ is a Born-layer observable, measured in jet cross sections and deep inelastic scattering. It describes the rendering at accessible wavelengths, not the $a \rightarrow 0$ limit.)

The Clay Millennium Prize problems on Yang–Mills existence and mass gap, and on Navier–Stokes existence and smoothness, are both problems about the continuum rendering, not about the physics. Lattice QCD already demonstrates the mass gap numerically at finite a —the gap is a property of the lattice Hamiltonian’s spectrum, visible and computable without any continuum limit. The Clay problem asks whether the rendering can be promoted to a rigorous free-standing continuum theory. This is a question in mathematics, not in physics: it asks whether the screen image can exist without the game engine that produces it. Similarly, the Navier–Stokes problem asks whether the continuum fluid equations have smooth solutions for all time—but Navier–Stokes is itself a long-wavelength rendering of molecular dynamics, and at the molecular scale there are no PDEs and no smoothness questions, just molecules colliding.

This does not mean the lattice Hamiltonian at spacing a is arbitrary. The requirement of rendering stability (1.11) constrains $H(a)$ severely: only a small family of Hamiltonians (parameterised by the renormalisable couplings) produce the same long-wavelength Born-layer physics. The constraints come from above (the rendered physics), not from below (a mythical sub-lattice structure).

The Einstein analogy. Einstein’s 1905 theory of Brownian motion is the cleanest historical illustration of the lattice-first viewpoint.

Einstein’s physics was about the molecules: discrete objects with a characteristic spacing (mean free path $\ell \sim 10^{-7}$ cm) and a characteristic timescale (collision time $\tau \sim 10^{-12}$ s). At the observational timescale of $\sim 10^{-2}$ s, the pollen grain’s trajectory looks continuous and random—a “Brownian path.” Wiener (1923) later gave this continuum picture a rigorous mathematical foundation: the Wiener measure on the space of continuous functions, producing paths that are continuous but nowhere differentiable.

The Wiener measure is the continuum rendering. It is completely blind to molecules. It knows nothing of collisions, of mean free paths, of the discrete impacts that drive the motion. It is a beautiful piece of mathematics, valid at timescales $t \gg \tau$, and it is not the physics. The physics is the molecules.

Now notice: you cannot push the Wiener measure down to 10^{-12} s. At that timescale, you see individual molecular collisions—discrete, ballistic, nothing like a continuous stochastic process. The nowhere-differentiable Wiener path is an artefact of the rendering; the actual molecular trajectory is piecewise smooth between collisions. The Wiener measure’s “roughness” is not a physical property of the pollen grain’s motion; it is what happens when you describe a discrete process with a continuum tool and push the tool past its domain of validity.

The analogy to quantum field theory is exact:

Brownian motion	Quantum field theory
Molecules (discrete substrate)	Lattice sites and links (Fock-layer engine)
Mean free path ℓ	Lattice spacing a
Pollen grain trajectory at $t \gg \tau$	Born-layer observables at $\lambda \gg a$
Wiener measure on continuous paths	Continuum path integral $\mathcal{D}\phi e^{iS}$
Rigorous construction of Wiener measure	Constructive QFT / Clay Yang–Mills problem
“What happens at 10^{-12} s?”	“What is the UV completion?”
Answer: molecules colliding	Answer: the lattice Hamiltonian; nothing deeper

Einstein never needed the Wiener measure. He needed the molecules. The Wiener measure came later, as a mathematical convenience for those who wanted to work with the rendering without thinking about the substrate. Demanding that the Wiener measure be rigorously defined before one can do physics would have inverted the intellectual priority entirely—it would have asked the rendering to validate the engine, rather than the other way around.

The same inversion has occurred in quantum field theory. The constructive QFT programme, the Clay problem, and the search for UV completions all treat the continuum rendering as the fundamental object and ask whether it can be made rigorous. The lattice-first viewpoint says: the engine is the lattice; the rendering is derived; and the rigour of the rendering is a mathematical question about the quality of an approximation, not a physical question about the foundations.

Relation to Wilson. This viewpoint is consistent with Kenneth Wilson’s understanding of the renormalisation group: quantum field theory is an effective description valid below a cutoff, and the RG describes how the effective parameters flow as the cutoff changes, with universality ensuring that different microscopic Hamiltonians can produce the same long-wavelength physics. Wilson’s lattice gauge theory (1974) was the first explicit construction of a gauge-invariant Fock-layer Hamiltonian on a discrete substrate, and it remains the most powerful non-perturbative tool in QCD.

Our viewpoint goes one step further in two directions. First, we insist on the ontological reading: the lattice is not a regularisation trick but the actual computational substrate of the Fock layer. The continuum limit is a mathematical convenience, not a physical destination. Second, we frame the whole picture in Fock-Born language: the bare parameters $\theta(a)$ are Fock-layer data; the rendered observables are Born-layer outputs; and renormalisability is the statement that the rendering is stable under re-meshing of the engine. The lattice spacing a is a parameter of the

game engine, free to vary; the physics on the screen is what the observer sees, and it does not depend on a at leading order.

A note on notation. Throughout this book we continue to write $\int d^3x$, $\partial/\partial x^\mu$, $\mathcal{D}\phi$, and other continuum symbols. These should be read as shorthand for their lattice counterparts: $\int d^3x \rightarrow a^3 \sum_{\mathbf{n}}$, $\partial/\partial x^\mu \rightarrow \Delta_\mu/a$, and $\mathcal{D}\phi \rightarrow \prod_{\mathbf{n}} d\phi_{\mathbf{n}}$. The continuum notation is compact and familiar; the lattice notation is what the Fock-layer engine actually computes.

Crucially, the limit $a/\lambda \rightarrow 0$ is taken *after* the observable has been calculated—that is, the limit is a Born-layer operation, not a Fock-layer one. The Fock layer always operates at finite a . One computes the Born-layer observable $\mathcal{O}_{\text{Born}}(\lambda; a)$ on the lattice, verifies that it is independent of a up to $o(a/\lambda)$ corrections (this is renormalisability), and reports the leading-order value as the physical prediction. The continuum limit is a statement about the quality of the rendering, not a requirement on the engine. The engine never takes a to zero; the Born layer does, as a mathematical convenience.

The diagnostic. Whenever a physical intuition narrates a sequence of events inside the Fock layer—particles emitted, paths taken, fields fluctuating, vacua teeming, barriers crossed—ask: is an observer submitting an observable and receiving an eigenvalue? If not, the narration is smuggling Born-layer language into the Fock layer.

The Fock layer has no events. It has one vector, rotating. Events belong to the Born layer, and they occur only when an observer interrogates the hidden state.

The textbook inversion and its intellectual history. How did the conflation become systematic? Through a pedagogical choice that hardened into a foundational commitment.

The historical sequence ran in the correct direction. Heisenberg (1925) and Schrödinger (1926) worked entirely with operators and wavefunctions. Dirac’s transformation theory (1927) established the operator framework as the foundation. Schwinger, Tomonaga, and Feynman independently solved renormalisation in QED around 1948–49. Of these three, Schwinger and Tomonaga worked within the operator formalism; Feynman introduced his diagrams and the path-integral picture. Dyson then proved the equivalence of Schwinger’s and Feynman’s approaches.

The two approaches are calculationally equivalent, but *pedagogically* they are not. The Feynman diagram comes with a built-in narrative: particles propagate, interact at vertices, and exchange other particles. This narrative is vivid and calculationally efficient. It is also, as we have catalogued above, a systematic Born-layer narration of Fock-layer bookkeeping. The operator approach carries no such narrative; it says: here is a Hamiltonian, here is a matrix element, compute it. The operator approach is harder to tell stories about, and stories are how physics gets taught.

Schwinger saw this clearly. He remarked that Feynman diagrams had brought “computation to the masses,” but insisted they were “pedagogy, not physics”—a calculational mnemonic, not an ontological picture. His source-theory program was an attempt to do all of quantum field theory without ever leaving the operator formalism. It was technically formidable, and it lost the pedagogical battle: the diagrams were simply easier to teach and to use.

Veltman, whose work with ’t Hooft established the renormalisability of Yang–Mills theory, was equally uneasy. In the opening of *Diagrammar* (1973), he and ’t Hooft wrote that the path integral “cannot replace the canonical formalism in defining a theory,” noting that path integrals in four dimensions are not well-defined without additional regularisation prescriptions that must be imported from the operator framework. The path integral borrows its meaning from the operator formalism; it does not supply its own.

Yet the textbooks inverted the priority. Textbook 1 opens with the path integral as Chapter I.2 and never looks back. Textbook 2 begins on page 1 with $\int \mathcal{D}\phi e^{iS}$. Textbook 3 starts canonical but abandon the operator formalism for the path integral by Part II. Textbook 4 labels the path integral the “modern approach” and the operator formalism the “old one.” The student absorbs a clear message: the Lagrangian path integral is the real quantum field theory; the Hamiltonian is a stepping stone you leave behind.

Once this ordering is established, every conflation in our catalogue follows naturally. If $\int \mathcal{D}\phi e^{iS}$ is the foundation, then $\phi(\mathbf{x}, t)$ at each time slice looks like the value the field “has,” and a sum over field configurations looks like a sum over histories the system “takes.” Virtual particles, vacuum fluctuations, exchange forces—these are all narrations of the path-integral bookkeeping as if it were a sequence of Born-layer events. The textbooks do not create the confusion maliciously; the confusion is built into the choice of starting point. Start from the path integral, and Born-layer language is the only language available to give the formalism physical meaning, because the Fock-layer structure has been suppressed.

The Fock-Born framework is, in a sense, the condensed-matter way of thinking exported to the whole of physics. In condensed matter, no one is confused about which of H and S is foundational. The rest of physics is catching up.

2 Dirac Notation, Basis Agnosticism, Direct Product, and Direct Sum

Before setting up the mathematics of quantum physics, we introduce the notation that will carry us through the book. Dirac's bra-ket notation is not merely a cosmetic choice — it encodes a genuinely useful idea about how vectors and their indices relate, and it connects naturally to the ML concepts of word embedding and position embedding.

2.1 The problem with dummy names

In ordinary linear algebra, a vector in $\mathbb{C}^n$ is written as a column of numbers, or given a name like $\mathbf{v}$. When we have a family of vectors, the vector may be indexed, e.g., $(\mathbf{v}_x)$ or $(\mathbf{v}_k)$. When the index is the natural label for the vector — when the whole point is that this vector corresponds to momentum $\mathbf{k}$, or position x , or a particle configuration $(n, \mathbf{k}_1, \dots, \mathbf{k}_n)$ — the dummy name $\mathbf{v}$ adds clutter without adding information.

Dirac's solution: write the vector as $|k\rangle$, $|x\rangle$, or $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$. The label inside the ket *is* the vector's name and its index simultaneously. No dummy letter required.

2.2 Kets, bras, and the inner product

A *ket* $|\psi\rangle$ is a vector in a complex Hilbert space $\mathcal{H}$. The corresponding *bra* $\langle\psi|$ is the dual vector: the conjugate-linear functional that maps any ket $|\phi\rangle$ to the complex number $\langle\psi|\phi\rangle$. In finite dimensions, if $|\psi\rangle$ is a column vector then $\langle\psi|$ is its conjugate transpose (a row vector).

The *inner product* of two kets $|\phi\rangle$ and $|\psi\rangle$ is written $\langle\phi|\psi\rangle$ and satisfies:

$$\langle\phi|\psi\rangle = \langle\psi|\phi\rangle^* \quad (\text{conjugate symmetry}), \quad (2.1)$$

$$\langle\psi|\alpha\phi_1 + \beta\phi_2\rangle = \alpha\langle\psi|\phi_1\rangle + \beta\langle\psi|\phi_2\rangle \quad (\text{linearity in the second argument}), \quad (2.2)$$

$$\langle\psi|\psi\rangle \geq 0, \quad \text{with equality iff } |\psi\rangle = 0 \quad (\text{positive definiteness}). \quad (2.3)$$

The norm of a ket is $\| |\psi\rangle \| = \sqrt{\langle\psi|\psi\rangle}$. Physical states are normalized: $\langle\psi|\psi\rangle = 1$. A linear operator A acts on kets from the left, $A|\psi\rangle$, and on bras from the right, $\langle\psi|A$; its matrix element between two states is the number $\langle\phi|A|\psi\rangle = \langle\phi|A\psi\rangle$.

A basis $\{|\lambda\rangle\}$ is *orthonormal* if:

$$\langle\lambda|\lambda'\rangle = \delta_{\lambda\lambda'} \quad (\text{discrete}), \quad \langle\lambda|\lambda'\rangle = \delta(\lambda - \lambda') \quad (\text{continuous}), \quad (2.4)$$

and *complete* (a resolution of the identity) if:

$$\sum_{\lambda} |\lambda\rangle \langle\lambda| = \mathbf{1} \quad (\text{discrete}), \quad \int d\lambda |\lambda\rangle \langle\lambda| = \mathbf{1} \quad (\text{continuous}). \quad (2.5)$$

Any ket $|\psi\rangle$ expands in an orthonormal basis as:

$$|\psi\rangle = \mathbf{1}|\psi\rangle = \sum_{\lambda} |\lambda\rangle \langle\lambda|\psi\rangle = \sum_{\lambda} c_{\lambda} |\lambda\rangle, \quad c_{\lambda} = \langle\lambda|\psi\rangle. \quad (2.6)$$

The coefficient $c_{\lambda} = \langle\lambda|\psi\rangle$ is the inner product of $|\psi\rangle$ with the basis vector $|\lambda\rangle$: it is the projection of $|\psi\rangle$ onto $|\lambda\rangle$, or equivalently the component of $|\psi\rangle$ in the λ direction. A linear operator A acts as:

$$A|\lambda\rangle = \sum_{\lambda'} |\lambda'\rangle \langle\lambda'|A|\lambda\rangle = \sum_{\lambda'} A_{\lambda'\lambda} |\lambda'\rangle, \quad A_{\lambda'\lambda} = \langle\lambda'|A|\lambda\rangle. \quad (2.7)$$

2.3 Dirac notation as matrix and vector notation

The abstract Dirac notation becomes completely concrete when we choose an orthonormal basis and label its vectors $|1\rangle, |2\rangle, \dots, |n\rangle$ (or $|i\rangle$ for short). Every ket becomes a column vector, every bra becomes a row vector, every operator becomes a matrix, and every abstract identity becomes a familiar linear algebra statement.

Basis vectors as standard basis vectors. The orthonormal basis $\{|i\rangle\}$ satisfies $\langle i|j\rangle = \delta_{ij}$. Concretely, in $\mathbb{C}^n$ we may take:

$$|1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad |2\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots \quad |n\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \quad (2.8)$$

The bra $\langle i| = |i\rangle^\dagger$ is the conjugate transpose: the row vector with a 1 in position i and zeros elsewhere.

A ket $|h\rangle$ is a column vector; $\langle i|h\rangle$ is its i -th element. Any state $|h\rangle \in \mathcal{H}$ is a column vector with complex entries. The inner product $\langle i|h\rangle$ picks out the i -th component:

$$\langle i|h\rangle = \langle i| |h\rangle = \begin{pmatrix} 0 & \dots & 1 & \dots & 0 \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_i \\ \vdots \\ h_n \end{pmatrix} = h_i. \quad (2.9)$$

The notation $\langle i|h\rangle = h_i$ is simply the inner product of the i -th standard basis vector with the column vector $|h\rangle$: it reads off the i -th component. The wavefunction $\psi(x) = \langle x|\psi\rangle$ is the continuum version of this: the x -th component of $|\psi\rangle$ in the position basis.

An operator A is a matrix; $\langle i|A|j\rangle$ is its (i, j) element. A linear operator $A : \mathcal{H} \rightarrow \mathcal{H}$ is represented in the basis $\{|i\rangle\}$ by the matrix with entries:

$$A_{ij} = \langle i|A|j\rangle. \quad (2.10)$$

This is the standard matrix element: A_{ij} is what A does to the j -th basis vector, projected onto the i -th basis vector. Concretely:

$$A = \begin{pmatrix} \langle 1|A|1\rangle & \langle 1|A|2\rangle & \dots \\ \langle 2|A|1\rangle & \langle 2|A|2\rangle & \dots \\ \vdots & & \ddots \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \dots \\ A_{21} & A_{22} & \dots \\ \vdots & & \ddots \end{pmatrix}. \quad (2.11)$$

The action of A on $|h\rangle$ in component form:

$$(A|h\rangle)_i = \langle i|Ah\rangle = \langle i|A|h\rangle = \langle i|A \sum_j h_j |j\rangle = \sum_j \langle i|A|j\rangle h_j = \sum_j A_{ij} h_j, \quad (2.12)$$

which is exactly the standard matrix-vector product $A_{ij}h_j$.

The resolution of the identity is the completeness of the basis. The completeness relation $\sum_i |i\rangle \langle i| = \mathbf{1}$ is the statement that $\{|i\rangle\}$ is a complete basis: every vector can be

expanded in it. In matrix language:

$$\sum_i |i\rangle \langle i| = \sum_i \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix} = \sum_i e_i e_i^T = I_n, \quad (2.13)$$

where e_i is the i -th standard basis vector and the sum of outer products $e_i e_i^T$ reconstructs the identity matrix. Applying to any $|h\rangle$:

$$\mathbf{1} |h\rangle = \left(\sum_i |i\rangle \langle i| \right) |h\rangle = \sum_i |i\rangle \underbrace{\langle i|h\rangle}_{h_i} = \sum_i h_i |i\rangle = |h\rangle. \quad (2.14)$$

The resolution of the identity simply says: any vector equals the sum of its components times the corresponding basis vectors. This is the completeness of the basis expansion, written in Dirac notation.

Inserting the resolution of the identity reads out matrix elements. The power of the resolution of the identity is that it can be inserted anywhere in a product of operators to expose the matrix elements. Consider the matrix product AB :

$$(AB)_{ij} = \langle i| AB |j\rangle = \langle i| A \underbrace{\left(\sum_k |k\rangle \langle k| \right)}_{\mathbf{1}} B |j\rangle = \sum_k \langle i| A |k\rangle \langle k| B |j\rangle = \sum_k A_{ik} B_{kj}. \quad (2.15)$$

This is matrix multiplication: $(AB)_{ij} = \sum_k A_{ik} B_{kj}$. Inserting the resolution of the identity has read out the matrix elements and performed the sum over the intermediate index k .

For a power of a matrix U^N :

$$(U^N)_{ij} = \langle i| U^N |j\rangle = \sum_{k_1, \dots, k_{N-1}} \langle i| U |k_1\rangle \langle k_1| U |k_2\rangle \cdots \langle k_{N-1}| U |j\rangle = \sum_{k_1, \dots, k_{N-1}} U_{i,k_1} U_{k_1,k_2} \cdots U_{k_{N-1},j}, \quad (2.16)$$

obtained by inserting $N - 1$ resolutions of the identity between the N factors of U . The sum over intermediate indices $(k_1, \dots, k_{N-1})$ is a sum over all sequences of intermediate basis states.

Time slicing in the path integral is matrix multiplication with intermediate basis insertions. This is exactly the time slicing of the path integral. The propagator:

$$K(x_f, T; x_i, 0) = \langle x_f | e^{-iHT} | x_i \rangle = (e^{-iHT})_{x_f, x_i} \quad (2.17)$$

is a matrix element of e^{-iHT} in the position basis. Time slicing writes $e^{-iHT} = (e^{-iH\varepsilon})^N$; inserting the position resolution of the identity $\mathbf{1} = \int dx |x\rangle \langle x|$ between each factor:

$$\begin{aligned} K(x_f, T; x_i, 0) &= \int dx_1 \cdots dx_{N-1} \langle x_f | e^{-iH\varepsilon} | x_{N-1} \rangle \cdots \langle x_1 | e^{-iH\varepsilon} | x_i \rangle \\ &= \int dx_1 \cdots dx_{N-1} (U_\varepsilon)_{x_f, x_{N-1}} \cdots (U_\varepsilon)_{x_1, x_i}, \end{aligned} \quad (2.18)$$

which is the matrix power (7.4) in the position basis, with the discrete sum over intermediate indices $k_1, \dots, k_{N-1}$ replaced by integrals over intermediate positions $x_1, \dots, x_{N-1}$. Each intermediate integration $\int dx_k$ is one insertion of the resolution of the identity; the product of matrix elements along a path $(x_i, x_1, \dots, x_{N-1}, x_f)$ is the amplitude of that path; and the full integral over all intermediate positions is the matrix power $(U_\varepsilon^N)_{x_f, x_i}$.

The path integral is thus demystified: it is the standard formula for the (x_f, x_i) matrix element of a matrix power U^N , written in the position basis, with the intermediate basis insertions made explicit. The sum over paths *is* the matrix multiplication; the path is the sequence of intermediate basis indices; and the amplitude of a path is the product of matrix elements along the sequence.

Dirac / abstract notation	Matrix / vector notation
$ h\rangle$	Column vector $(h_1, \dots, h_n)^T$
$\langle h $	Row vector $(h_1^*, \dots, h_n^*)$
$\langle i h\rangle$	i -th component h_i of $ h\rangle$
$\langle i A j\rangle$	Matrix element A_{ij}
$\sum_i i\rangle \langle i = \mathbf{1}$	$\sum_i e_i e_i^T = I_n$ (completeness)
$ h\rangle = \sum_i i\rangle \langle i h\rangle$	$\mathbf{h} = \sum_i h_i e_i$ (basis expansion)
$\langle i AB j\rangle = \sum_k A_{ik} B_{kj}$	Matrix multiplication $(AB)_{ij}$
Inserting $\sum_k k\rangle \langle k $	Summing over intermediate index k
Path integral $\int dx_1 \cdots dx_{N-1}$	Sum over intermediate indices in U^N
Path amplitude $\prod_k (U_\varepsilon)_{x_{k+1}, x_k}$	Product of matrix elements along path

2.4 One-hot vectors, dense vectors, and the continuous index

In the previous subsection, every basis vector $|i\rangle$ was a one-hot column vector: a single 1 in position i and zeros elsewhere. This is a specific representation, natural for computation but not the only one. We now make three observations that open the notation to its full generality.

One-hot representation

The basis vectors $|1\rangle, |2\rangle, \dots, |n\rangle$ in the one-hot representation are:

$$(|i\rangle)_j = \delta_{ij} = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases} \quad (2.19)$$

This is the standard basis of $\mathbb{C}^n$: each basis vector has all its weight on a single index. The index i labels *which* basis vector; the entry j labels the *component*. The one-hot vector says: “I am entirely the i -th basis vector and nothing else.”

A general state $|h\rangle$ in this basis is a column of complex numbers:

$$|h\rangle = \sum_i h_i |i\rangle, \quad h_i = \langle i|h\rangle. \quad (2.20)$$

The components h_i are the amplitudes on each basis vector. This is the *dense* representation of $|h\rangle$: a vector with (in general) nonzero entries everywhere.

Dense representation: when the basis vector itself is a superposition

Now consider a different orthonormal basis $\{|\alpha\rangle\}$ of the same Hilbert space. In the one-hot basis $\{|i\rangle\}$, each $|\alpha\rangle$ is a dense vector:

$$(|\alpha\rangle)_i = \langle i|\alpha\rangle = U_{i\alpha}, \quad (2.21)$$

where $U_{i\alpha}$ is the unitary change-of-basis matrix. The vector $|\alpha\rangle$ is one-hot in its *own* basis (it is the α -th standard basis vector of $\{|\alpha\rangle\}$) but dense in the $\{|i\rangle\}$ basis (it is a superposition $\sum_i U_{i\alpha} |i\rangle$).

A general state can be expanded in either basis:

$$|h\rangle = \sum_i h_i |i\rangle = \sum_\alpha \tilde{h}_\alpha |\alpha\rangle, \quad h_i = \langle i|h\rangle, \quad \tilde{h}_\alpha = \langle \alpha|h\rangle. \quad (2.22)$$

The physical vector $|h\rangle$ is the same; the components h_i and $\tilde{h}_\alpha$ are its representations in two different bases. In one basis a vector may be one-hot (a single basis state); in another it is dense (a superposition). Dirac notation treats both on equal footing: $\langle i|h\rangle$ and $\langle \alpha|h\rangle$ are both inner products, both read off components, and both follow the same rules.

In ML terms: the one-hot representation is the token index; the dense representation is the embedding vector. The embedding matrix U converts one to the other, exactly as the change-of-basis matrix does here.

Continuous index: the position basis $|x\rangle$

So far the index i has been discrete: $i = 1, 2, \dots, n$. But in quantum mechanics, many physical observables—position, momentum, field amplitude—have a continuous spectrum. The index then becomes a real number $x \in \mathbb{R}$.

The basis vector $|x\rangle$ can still be thought of as an infinite-dimensional vector, with x labelling the entries. The “component” of $|x\rangle$ at position x' is:

$$(|x\rangle)_{x'} = \langle x'|x\rangle = \delta(x' - x), \quad (2.23)$$

where $\delta(x' - x)$ is the Dirac delta function: zero everywhere except at $x' = x$, where it is infinite, with $\int dx' \delta(x' - x) = 1$. This is the continuous analogue of the Kronecker delta δ_{ij} : the one-hot vector generalised to a continuous index.

	Discrete index i	Continuous index x
Basis vector	$ i\rangle$	$ x\rangle$
One-hot / delta	$(i\rangle)_j = \delta_{ij}$	$\langle x' x\rangle = \delta(x' - x)$
Orthonormality	$\langle i j\rangle = \delta_{ij}$	$\langle x x'\rangle = \delta(x - x')$
Completeness	$\sum_i i\rangle \langle i = \mathbf{1}$	$\int dx x\rangle \langle x = \mathbf{1}$
Component of $ h\rangle$	$h_i = \langle i h\rangle$	$\psi(x) = \langle x h\rangle$
Expansion	$ h\rangle = \sum_i h_i i\rangle$	$ h\rangle = \int dx \psi(x) x\rangle$
normalization	$\sum_i h_i ^2 = 1$	$\int dx \psi(x) ^2 = 1$
Matrix element	$A_{ij} = \langle i A j\rangle$	$A(x, x') = \langle x A x'\rangle$
Matrix product	$\sum_k A_{ik} B_{kj}$	$\int dx'' A(x, x'') B(x'', x')$

Every formula from the discrete case has a continuous counterpart obtained by two substitutions: $\delta_{ij} \rightarrow \delta(x - x')$ and $\sum_i \rightarrow \int dx$. The Dirac notation $\langle x|h\rangle$, $\langle x|A|x'\rangle$, $\int dx |x\rangle \langle x| = \mathbf{1}$ is identical in form to $\langle i|h\rangle$, $\langle i|A|j\rangle$, $\sum_i |i\rangle \langle i| = \mathbf{1}$. The notation does not care whether the index is discrete or continuous.

The wavefunction is a dense representation

The wavefunction $\psi(x) = \langle x|h\rangle$ is the dense representation of the state $|h\rangle$ in the position basis. It is the continuous analogue of the component vector $h_i = \langle i|h\rangle$: a complex-valued function that gives the amplitude at each point x .

Just as a dense vector h_i in the discrete case is a superposition $|h\rangle = \sum_i h_i |i\rangle$ of one-hot basis vectors, the wavefunction is a superposition of Dirac deltas:

$$|h\rangle = \int dx \psi(x) |x\rangle. \quad (2.24)$$

Each $|x\rangle$ is a “one-hot” state—perfectly localised at position x —and $\psi(x)$ is the amplitude for that location. A smooth wavefunction is a dense vector in the position basis: amplitude spread across many positions, just as a dense vector has entries spread across many indices.

Conversely, $|h\rangle$ can be expanded in the momentum basis $\{|p\rangle\}$:

$$|h\rangle = \int \frac{dp}{2\pi\hbar} \tilde{\psi}(p) |p\rangle, \quad \tilde{\psi}(p) = \langle p|h\rangle. \quad (2.25)$$

The momentum-space wavefunction $\tilde{\psi}(p)$ is the dense representation in the momentum basis. A state localised in position ($\psi(x) \approx \delta(x - x_0)$, one-hot in x) is delocalised in momentum ($\tilde{\psi}(p)$ is spread over all p , dense in p), and vice versa. This is the uncertainty principle, expressed as a statement about representations: a vector that is one-hot in one basis is necessarily dense in a complementary basis.

The change-of-basis matrix between position and momentum is the Fourier kernel:

$$\langle x|p\rangle = e^{ipx/\hbar}, \quad \tilde{\psi}(p) = \int dx e^{-ipx/\hbar} \psi(x). \quad (2.26)$$

This is the continuous analogue of the unitary matrix $U_{i\alpha}$ in (2.21): it converts between two dense representations of the same physical vector.

Summary: three levels of generality

Level	Index	Basis vector	Dense vector
Finite discrete	$i = 1, \dots, n$	$(i\rangle)_j = \delta_{ij}$ (one-hot)	$h_i = \langle i h\rangle$ (components)
Countable discrete	$n = 0, 1, 2, \dots$ (Fock)	$(n\rangle)_m = \delta_{nm}$ (one-hot)	$c_n = \langle n h\rangle$ (Fock amplitudes)
Continuous	$x \in \mathbb{R}$	$\langle x' x\rangle = \delta(x' - x)$ (delta)	$\psi(x) = \langle x h\rangle$ (wavefunction)

At every level, the structure is the same: an orthonormal basis (one-hot or delta), a completeness relation (sum or integral), and a dense representation (components or wavefunction). Dirac notation unifies all three levels with a single set of symbols: $|\cdot\rangle$, $\langle\cdot|$, $\langle\cdot|\cdot\rangle$. Whether the index is finite, countable, or continuous; whether the representation is one-hot or dense; whether the basis is position, momentum, energy, or Fock number—the notation is the same, and the rules are the same. This is why we use it throughout this book.

2.5 The wavefunction is a coefficient, not a function

The most important consequence of (2.6) is a reinterpretation of the wavefunction. In introductory quantum mechanics, the wavefunction $\psi(x)$ is often introduced as a fundamental object: a complex-valued function of position, satisfying the Schrödinger equation. The Dirac notation reveals what it actually is: a *coefficient* of the state $|\psi\rangle$ in the position basis $\{|x\rangle\}$.

The position basis vectors $|x\rangle$ ($x \in \mathbb{R}$) are an orthonormal, complete family:

$$\langle x|x'\rangle = \delta(x - x'), \quad \int dx |x\rangle \langle x| = \mathbf{1}. \quad (2.27)$$

Inserting the resolution of the identity into any state $|\psi\rangle$:

$$|\psi\rangle = \int dx |x\rangle \underbrace{\langle x|\psi\rangle}_{\psi(x)} = \int dx \psi(x) |x\rangle. \quad (2.28)$$

The wavefunction $\psi(x) = \langle x|\psi\rangle$ is the coefficient of $|\psi\rangle$ along the basis vector $|x\rangle$. It is not the state itself; it is the state's representation in the position basis. The state $|\psi\rangle$ exists independently of this representation — it is an abstract vector in $\mathcal{H}$ — and the wavefunction is what you get when you choose to read it off in the position basis.

This is the same as the relationship between a vector $\mathbf{v} \in \mathbb{R}^3$ and its components (v_x, v_y, v_z) in a chosen coordinate system: the vector is primary; the components depend on the choice of axes. Different choices of basis give different “wavefunctions” for the same state $|\psi\rangle$.

The momentum wavefunction. The momentum basis vectors $|p\rangle$ are defined by the eigenvalue equation $\hat{p}|p\rangle = p|p\rangle$ and satisfy:

$$\langle p|p'\rangle = \delta(p - p'), \quad \int dp |p\rangle \langle p| = \mathbf{1}. \quad (2.29)$$

The momentum wavefunction $\tilde{\psi}(p) = \langle p|\psi\rangle$ is the coefficient of $|\psi\rangle$ in the momentum basis.

Deriving $\langle x|p\rangle$ from the commutation relation. The position-basis representation of a momentum eigenstate is a plane wave. We derive this from the commutation relation $[\hat{x}, \hat{p}] = i$ alone.

The momentum eigenvalue equation $\hat{p}|p\rangle = p|p\rangle$ says that $|p\rangle$ is a state with definite momentum p . To find its position-basis representation $f(x) = \langle x|p\rangle$, we sandwich both sides between $\langle x|$ and $|p\rangle$:

$$\langle x|\hat{p}|p\rangle = p \langle x|p\rangle = p f(x). \quad (2.30)$$

On the left, we need $\langle x|\hat{p}$. In the position basis, $\hat{p}$ acts as $-i d/dx$:

$$\langle x|\hat{p} = -i \frac{d}{dx} \langle x|. \quad (2.31)$$

(This is derived in Section 4 by verifying $[\hat{x}, \hat{p}] = i$.) Therefore:

$$\langle x|\hat{p}|p\rangle = -i \frac{d}{dx} \langle x|p\rangle = -i \frac{df}{dx}. \quad (2.32)$$

Equating the two expressions:

$$-i \frac{df}{dx} = p f(x). \quad (2.33)$$

This is a first-order ODE with solution:

$$f(x) = C e^{ipx}, \quad (2.34)$$

where C is a normalization constant. To fix C , impose $\langle p|p'\rangle = \delta(p - p')$:

$$\langle p|p'\rangle = \int dx \langle p|x\rangle \langle x|p'\rangle = |C|^2 \int dx e^{i(p'-p)x} = |C|^2 \cdot 2\pi \delta(p' - p). \quad (2.35)$$

Setting this equal to $\delta(p - p')$ requires $|C|^2 = 1/(2\pi)$, so $C = 1/\sqrt{2\pi}$ (choosing the conventional positive real phase). The result:

$$\boxed{\langle x|p\rangle = \frac{e^{ipx}}{\sqrt{2\pi}}.} \quad (2.36)$$

This is the plane wave: the position-basis representation of a definite-momentum state. It was not postulated; it was derived from the single algebraic relation $[\hat{x}, \hat{p}] = i$. The entire Fourier transform machinery of quantum mechanics follows from this one commutator.

The Fourier transform as a change of basis. The change of basis from position to momentum wavefunctions follows by inserting a resolution of the identity:

$$\tilde{\psi}(p) = \langle p | \psi \rangle = \int dx \langle p | x \rangle \langle x | \psi \rangle = \int dx \frac{e^{-ipx}}{\sqrt{2\pi}} \psi(x), \quad (2.37)$$

which is precisely the Fourier transform of $\psi(x)$. The factor $\langle p | x \rangle = \langle x | p \rangle^* = e^{-ipx}/\sqrt{2\pi}$ is the change-of-basis kernel—the continuous analogue of the unitary matrix $U_{i\alpha} = \langle i | \alpha \rangle$ that relates two discrete bases.

The momentum basis $|p\rangle$ can be expanded in the position basis using the completeness relation $\mathbf{1} = \int dx |x\rangle \langle x|$:

$$|p\rangle = \int dx |x\rangle \langle x | p \rangle = \int dx \frac{e^{ipx}}{\sqrt{2\pi}} |x\rangle, \quad (2.38)$$

a superposition of position eigenstates with sinusoidal coefficients: the plane wave. Choosing the momentum basis is choosing to expand the state in plane waves; choosing the position basis is choosing to expand in delta functions. The state $|\psi\rangle$ is the same object in both cases.

In ML terms: the Fourier transform is a fixed, analytically known change-of-basis matrix between two representations of the same data. The position representation $\psi(x)$ is like a pixel-space image; the momentum representation $\tilde{\psi}(p)$ is like its frequency-space DCT. Both encode the same information; the Fourier transform converts between them; and neither is more fundamental than the other.

2.6 Word embedding and position embedding in transformers

The Dirac ket $|\lambda\rangle$ is, in ML language, an *embedding* of the discrete or continuous label λ into a vector space. The connection is exact.

Word embedding. In a transformer, each token from a vocabulary of size V is embedded as a vector in $\mathbb{R}^d$: the embedding table $E \in \mathbb{R}^{V \times d}$ maps token index w to the vector $E[w] \in \mathbb{R}^d$. In Dirac notation: $E[w] = |w\rangle$. The embedding of the word “quantum” is $|\text{“quantum”}\rangle$. The inner product $\langle \text{“red”} | \text{“blue”} \rangle$ measures the cosine similarity between the two word vectors. The Dirac notation makes the key insight explicit: the word “quantum” *is* the index of a vector; the vector $|\text{“quantum”}\rangle$ *is* its embedding. No dummy letter needed.

Position embedding. In a transformer, position $t \in \{1, \dots, T\}$ is embedded as a vector $p_t \in \mathbb{R}^d$. In Dirac notation: $p_t = |t\rangle$. The sinusoidal position embedding is one specific choice of basis for $|t\rangle$. This is exactly the position vector used in quantum mechanics: $|x\rangle$ is the position eigenstate, a vector in $L^2(\mathbb{R})$ indexed by position x . The transformer’s position embedding $|t\rangle$ is the finite-dimensional, discrete analogue of the quantum position state $|x\rangle$.

Momentum basis as sinusoidal embedding. Just as a transformer can use sinusoidal position embeddings, quantum mechanics uses the momentum basis $|p\rangle$ — the Fourier dual of the position basis, given by (2.38): a superposition of position eigenstates with sinusoidal weights $e^{ipx}/\sqrt{2\pi}$. The choice between $|x\rangle$ and $|p\rangle$ is a choice of basis, related by the Fourier transform (5.13) — exactly the freedom of choosing between learned and sinusoidal position embeddings in a transformer.

2.7 The power of Dirac notation: basis agnosticism

Everything in the previous subsections was written using one-hot basis vectors $|1\rangle, |2\rangle, \dots, |n\rangle$: column vectors with a single 1 and the rest zeros. But this is only one possible basis. The deepest feature of Dirac notation is that *every result holds in any orthonormal basis*, and the notation never commits to a specific one unless forced.

This is not a cosmetic point. It is the key to understanding how the Born layer works: the observer *chooses* which basis to read the hidden state in, and the choice determines what kind of classical output appears. Basis agnosticism is the mathematical expression of the Born layer’s multi-lingual capability.

A one-hot ket in one basis is a dense vector in another

Let $\{|i\rangle\}$ and $\{|\alpha\rangle\}$ be two orthonormal bases of the same Hilbert space. In the $\{|i\rangle\}$ basis, the vector $|\alpha\rangle$ is a column vector with components $\langle i|\alpha\rangle$:

$$|\alpha\rangle = \sum_i |i\rangle \langle i|\alpha\rangle = \sum_i \langle i|\alpha\rangle |i\rangle. \quad (2.39)$$

In the $\{|\alpha\rangle\}$ basis, this same vector is one-hot: a single 1 with the rest zeros. In the $\{|i\rangle\}$ basis, it is a dense vector—a superposition with (in general) nonzero entries everywhere. The physical vector has not changed; only the basis has.

The same state $|h\rangle$ has different component representations in different bases:

$$|h\rangle = \sum_i h_i |i\rangle = \sum_\alpha \tilde{h}_\alpha |\alpha\rangle, \quad h_i = \langle i|h\rangle, \quad \tilde{h}_\alpha = \langle \alpha|h\rangle. \quad (2.40)$$

In one basis a state may be one-hot (a single definite value); in another it is dense (a superposition spread across many values). Dirac notation treats both on equal footing: $\langle i|h\rangle$ and $\langle \alpha|h\rangle$ are both inner products, both read off components, and both follow the same rules.

Every formula holds in every basis

Every formula in the previous subsections—inner products, matrix elements, the resolution of the identity, matrix multiplication, the path integral—was derived using $|i\rangle$ as a one-hot vector. But the derivations used only two properties:

$$\langle i|j\rangle = \delta_{ij} \quad (\text{orthonormality}), \quad \sum_i |i\rangle \langle i| = \mathbf{1} \quad (\text{completeness}). \quad (2.41)$$

Any orthonormal basis satisfies both. Therefore every result holds with $|i\rangle$ replaced by $|\alpha\rangle, |x\rangle, |n\rangle, |p\rangle, |E\rangle$, or any other orthonormal basis—without changing a single equation.

The symbol $|\alpha\rangle$ does not specify whether it is one-hot or a dense superposition. It does not need to. The formalism works the same way in both cases, because it depends only on orthonormality and completeness, not on the specific components.

An observable defines its own basis

This is the central connection between basis agnosticism and the Born layer.

Every observable $\hat{O}$ is a Hermitian operator. The spectral theorem says it has real eigenvalues $\{\lambda\}$ and orthonormal eigenvectors $\{|\lambda\rangle\}$ satisfying $\hat{O}|\lambda\rangle = \lambda|\lambda\rangle$. The observable *defines its own basis*: the eigenbasis $\{|\lambda\rangle\}$, with each eigenvector named by its eigenvalue. The eigenvalue is the classical label—the “word” in the Born-layer vocabulary—and the eigenvector is the direction in Hilbert space that corresponds to that word.

Different observables define different bases:

Observable $\hat{\mathcal{O}}$	Eigenbasis $\{ \lambda\rangle\}$	Eigenvalues	Vocabulary
Position $\hat{x}$	$\{ x\rangle\}$ (delta functions)	$x \in \mathbb{R}$	“particle is at x ”
Momentum $\hat{p}$	$\{ p\rangle\}$ (plane waves)	$p \in \mathbb{R}$	“particle has momentum p ”
Energy H	$\{ E_n\rangle\}$ (energy eigenstates)	E_n	“system has energy E_n ”
Spin- z $\hat{S}_z$	$\{ \uparrow\rangle, \downarrow\rangle\}$	$\pm\hbar/2$	“spin up” or “spin down”
Particle number $\hat{n}$	$\{ n\rangle\}$ (Fock states)	$n = 0, 1, 2, \dots$	“ n particles present”

The hidden state $|h_t\rangle$ is the same vector in every case. The observable determines which basis to use for the readout. This is what “multi-lingual” means: the same hidden state, read in the position language, gives position probabilities; read in the momentum language, gives momentum probabilities; read in the energy language, gives energy probabilities. The observer’s choice of observable selects the language.

The Born rule is the readout mechanism:

$$P(\lambda) = |\langle\lambda|h_t\rangle|^2 = |\text{component of } h_t \text{ along } |\lambda\rangle|^2. \quad (2.42)$$

The inner product $\langle\lambda|h_t\rangle$ is the projection of the hidden state onto the eigenvector named λ —the “attention score” between the hidden state and the query vector $|\lambda\rangle$. The squared modulus gives the probability. Different observables provide different sets of query vectors $\{|\lambda\rangle\}$, and the same hidden state h_t returns different probability distributions depending on which set is used.

In ML terms: the observable is the *read head* or *unembedding matrix*. Each eigenvector $|\lambda\rangle$ is a row of the unembedding matrix; the inner product $\langle\lambda|h_t\rangle$ is one entry of the logit vector; and the Born rule $|\langle\lambda|h_t\rangle|^2$ is the softmax output (with the squared modulus playing the role of the softmax exponential, but with automatic normalisation from the completeness relation).

ML (transformer)	QM (Fock-Born)
Residual stream $r \in \mathbb{R}^d$	Hidden state $h_t \in \mathcal{H}$
Unembedding matrix $W_U \in \mathbb{R}^{V \times d}$	Observable’s eigenbasis $\{ \lambda\rangle\}$
Logit $\ell_w = W_U[w] \cdot r$	Amplitude $c_\lambda = \langle\lambda h_t\rangle$
Token vocabulary $\{w\}$	Eigenvalue vocabulary $\{\lambda\}$
$\text{softmax}(\ell)_w$	Born probability $ c_\lambda ^2$
Choice of read head	Choice of observable

The key difference: in a transformer, there is typically one read head (the unembedding matrix is fixed). In quantum mechanics, the observer can submit *any* Hermitian operator as the observable, and each one defines a different read head with a different vocabulary. The Born layer is interactive: the observer chooses the language at measurement time, not at training time.

What matters is the inner products

A vector $|h\rangle$ exists independently of any basis. Its components $\langle\alpha|h\rangle$ depend on the chosen basis $\{|\alpha\rangle\}$, but the vector itself does not. All physical predictions are inner products: the Born rule gives $P(\lambda) = |\langle\lambda|h\rangle|^2$; the expectation value gives $\langle\mathcal{O}\rangle = \langle h|\mathcal{O}|h\rangle$; the transition amplitude

is $\langle f|i \rangle$. What matters is not the components in any particular basis but the inner products between vectors.

The position wavefunction $\psi(x) = \langle x|h \rangle$, the momentum wavefunction $\tilde{\psi}(p) = \langle p|h \rangle$, the energy representation $c_n = \langle E_n|h \rangle$, and the Fock representation $c_{\{\mathbf{n}_k\}} = \langle \{\mathbf{n}_k\}|h \rangle$ are all encodings of the same state $|h \rangle$. The Schrödinger equation $\dot{h}_t = -iHh_t$ is a statement about the vector $h_t \in \mathcal{H}$, not about its components. Writing it in the position basis gives the Schrödinger wave equation $i\hbar\partial_t\psi(x,t) = \hat{H}\psi(x,t)$; writing it in the momentum basis gives $i\hbar\partial_t\tilde{\psi}(p,t) = \hat{H}\tilde{\psi}(p,t)$; writing it in the energy basis gives $\dot{c}_n = -iE_nc_n/\hbar$. All are the same equation in different clothes.

In ML terms: two different transformer models may learn completely different embedding matrices, but if the inner products between semantically related words are preserved, the models compute the same downstream tasks. The embedding is a choice of basis for the semantic space; the semantic relations are the inner products.

The essential structure, maximally compressed

Basis agnosticism is not a convenience—it is what makes Dirac notation the right language for quantum mechanics. The physically meaningful statements are:

- The hidden state h_t is a unit vector in a Hilbert space.
- The Schrödinger equation $\dot{h}_t = -iHh_t$ rotates it.
- An observable $\hat{O}$ defines a basis $\{|\lambda\rangle\}$ (its eigenvectors, named by eigenvalues).
- The Born probability of eigenvalue λ is the squared length of the projection of h_t onto the eigensubspace of λ : $P(\lambda) = |\langle \lambda|h_t \rangle|^2$.
- The Bohr collapse keeps the projected superposition.

None of these statements refers to a specific basis. The Dirac notation $|h_t\rangle$, $\langle \lambda|$, $P(\lambda) = \|P_\lambda h_t\|^2$ expresses them without choosing coordinates, just as a vector equation $\mathbf{F} = m\mathbf{a}$ expresses Newton’s law without choosing an axis system.

This philosophy is not unique to quantum mechanics. Tensor calculus writes the laws of general relativity as $G_{\mu\nu} = 8\pi GT_{\mu\nu}$ —valid in every coordinate system, committing to none. Differential forms write Maxwell’s equations as $dF = 0$, $d*F = J$ —valid in every frame, with the basis agnosticism built into the notation itself. In each case, the notation is designed to express the coordinate-free structure, and the specific components emerge only when a basis is chosen for a particular calculation.

Dirac notation does the same for quantum mechanics: it expresses the Hilbert space structure—inner products, operators, completeness—without committing to a basis. The one-hot representation is a tool for concrete calculation. The Dirac notation is the language of the physics.

Foreshadowing: Dirac’s transformation theory

The basis agnosticism of Dirac notation is not merely a notational preference—it is the foundation of Dirac’s *transformation theory* (1927), one of the great unifications of quantum mechanics. Dirac showed that the Schrödinger wave mechanics (formulated in the position basis $\{|x\rangle\}$), the Heisenberg matrix mechanics (formulated in the energy basis $\{|E_n\rangle\}$), and any other basis representation are related by unitary transformations—changes of basis in the Hilbert space. The physics is invariant under these transformations; only the representation changes.

In Dirac’s language: the physical content of quantum mechanics is the set of inner products $\langle \phi|\psi \rangle$ and $\langle \phi|A|\psi \rangle$ between states and operators. A change of basis $|i\rangle \rightarrow |\alpha\rangle = \sum_i U_{i\alpha} |i\rangle$ is a unitary transformation U that changes the components of every vector and every matrix but

leaves all inner products invariant:

$$\langle \phi | \psi \rangle = \sum_i \phi_i^* \psi_i = \sum_\alpha \tilde{\phi}_\alpha^* \tilde{\psi}_\alpha. \quad (2.43)$$

This is the quantum analogue of the invariance of the spacetime interval under Lorentz transformations: the coordinates change, but the inner product is preserved.

Transformation theory goes further: every symmetry of the Hamiltonian corresponds to a unitary transformation that commutes with H , and the classification of quantum systems by their symmetries is the classification of the representations of these transformations. We develop this systematically later.

2.8 Fock basis vectors as embeddings of world states

The Fock basis vector:

$$|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle \quad (2.44)$$

is an embedding of a complete physical world state: a universe containing exactly n particles with momenta $\mathbf{k}_1, \dots, \mathbf{k}_n$. The label inside the ket is the full specification of the world state; the ket is its vector representation in Fock space. The Fock hidden state:

$$h_t = \sum_{n, \mathbf{k}} c(n, \mathbf{k}; t) |n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle \quad (2.45)$$

is an expansion in Fock basis vectors, with coefficients $c(n, \mathbf{k}; t) = \langle n; \mathbf{k}_1, \dots, \mathbf{k}_n | h_t$ playing the role of the wavefunction in the Fock basis. The Schrödinger equation rotates h_t within Fock space; the Born rule computes the inner product $|\langle o | h_t \rangle|^2$ between h_t and a basis state $|o\rangle$ to give the probability of observing configuration o .

The analogy with CLIP (Contrastive Language-Image Pretraining) is instructive. CLIP learns a joint embedding of images and text such that $\langle \text{image} | \text{caption} \rangle$ is high when the image matches the caption. The Fock basis vector $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ plays the same role: it embeds an entire physical configuration into Hilbert space. The Born rule computes the inner product between the current world state h_t and the observed configuration $|o\rangle$, giving the probability of that observation.

Dirac notation	ML concept	Example
$ \lambda\rangle$	Embedding of label λ	$ \mathbf{k}\rangle, x\rangle, \text{word}\rangle$
$\langle \psi $	Dual vector (row vector)	Attention query
$\langle \lambda \psi \rangle$	Component / projection	Wavefunction $\psi(\lambda)$, attention score
$\sum_\lambda c_\lambda \lambda\rangle$	Weighted superposition	Residual stream, quantum state
$A \lambda\rangle$	Linear map / weight matrix	Attention, Hamiltonian
Basis $\{ x\rangle\}$	Position embedding	Position eigenstates
Basis $\{ p\rangle\}$	Sinusoidal embedding	Fourier / momentum basis
Fourier transform	Change of basis	Position $\leftrightarrow$ momentum
$ \langle o h \rangle ^2$	Softmax / classifier output	Born rule
$ n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$	CLIP embedding of world state	Fock basis vector

The Dirac ket is the notation that makes the following statement natural: the universe maintains a unit vector h_t in Fock space, and the wavefunction in any basis is the collection of inner

products of h_t with the basis vectors. The position wavefunction $\psi(x, t) = \langle x | h_t \rangle$, the momentum wavefunction $\tilde{\psi}(p, t) = \langle p | h_t \rangle$, and the Fock wavefunction $c(n, \mathbf{k}; t) = \langle n; \mathbf{k} | h_t \rangle$ are all the same state, read off in different bases. The Schrödinger equation rotates the vector; the basis choice determines how the rotation looks as a differential equation.

2.9 Superposition: the hidden state as a weighted sum of basis vectors

The expansion (2.6) is not merely a notational convenience. It encodes a profound physical fact: the hidden state $h_t = |\psi(t)\rangle$ is a *superposition* — a weighted sum of basis vectors, with complex coefficients, existing simultaneously in all basis directions at once. The superposition is not a statement about our ignorance of which basis state the system “really” is in. It is the complete description of the system. The system is in all basis states simultaneously, with amplitudes $c_\lambda = \langle \lambda | \psi \rangle$ that determine the Born probabilities when a measurement is eventually made.

Superposition in any basis. Given an orthonormal basis $\{|\lambda\rangle\}$ of $\mathcal{H}$, any state $|\psi\rangle$ expands as:

$$|\psi\rangle = \sum_{\lambda} c_{\lambda} |\lambda\rangle, \quad c_{\lambda} = \langle \lambda | \psi \rangle, \quad \sum_{\lambda} |c_{\lambda}|^2 = 1. \quad (2.46)$$

For a continuous basis such as the position eigenstates $\{|x\rangle\}$, the sum becomes an integral:

$$|\psi\rangle = \int dx \, c(x) |x\rangle, \quad c(x) = \langle x | \psi \rangle = \psi(x), \quad \int |c(x)|^2 dx = 1. \quad (2.47)$$

The wavefunction $\psi(x) = \langle x | \psi \rangle$ is the coefficient of the superposition in the position basis. The momentum wavefunction $\tilde{\psi}(p) = \langle p | \psi \rangle$ is the coefficient in the momentum basis. Both expansions describe the same $|\psi\rangle$; the Fourier transform converts between them. The hidden state is *basis-agnostic*: it exists independently of any expansion, and can be read off in any basis by taking the appropriate inner products.

The Fock hidden state as a superposition of world configurations. In quantum field theory, the Fock hidden state:

$$h_t = \sum_{n=0}^{\infty} \int \prod_{k=1}^n \frac{d^3 \mathbf{k}_k}{(2\pi)^3} c(n; \mathbf{k}_1, \dots, \mathbf{k}_n; t) |n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle \quad (2.48)$$

is a superposition over all possible world configurations: the vacuum, all one-particle states, all two-particle states, and so on. Each basis vector $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ is *monosemantic*: it encodes an exact, unambiguous physical configuration — precisely n particles with precisely these momenta and quantum numbers. The Fock hidden state h_t is *polysemantic*: it is a superposition of all these monosemantic basis vectors simultaneously, with complex amplitudes $c(n; \mathbf{k}_1, \dots; t)$ that the Schrödinger RNN rotates over time. The Born rendering layer projects h_t onto a chosen basis and produces one classical outcome; the superposition itself is never directly observed.

Connection to the residual stream of a transformer. The hidden state h_t of quantum mechanics and the residual stream of a large language model are the same kind of object: a dense vector in a high-dimensional space, a superposition of many features simultaneously.

In a transformer, the residual stream at each layer is a vector $r \in \mathbb{R}^d$ of dimension $d \sim 10^3$ – 10^4 . Sparse autoencoder analyses (such as those applied to GPT-class models by researchers studying mechanistic interpretability) decompose this vector as a superposition of *monosemantic basis vectors* drawn from a large learned dictionary $\{\mathbf{e}_i\}_{i=1}^D$ (with $D \gg d$, typically $D \sim 10^4$ – 10^6):

$$r \approx \sum_{i \in S} a_i \mathbf{e}_i, \quad (2.49)$$

where S is a small sparse subset of active features and each $\mathbf{e}_i$ is a monosemantic direction in $\mathbb{R}^d$ corresponding to a specific, interpretable concept (“Paris”, “the word ‘not’”, “a Python function”, etc.). The residual stream r is *polysemantic*: it encodes many concepts simultaneously in a single dense vector; the sparse autoencoder expands it as a superposition of monosemantic directions.

The parallel with quantum superposition is direct:

Quantum mechanics / QFT	Transformer residual stream
Fock basis vector $ n; \mathbf{k}_1, \dots\rangle$	Monosemantic SAE feature $\mathbf{e}_i$ (one concept, one configuration)
Fock hidden state h_t (polysemantic)	Residual stream r (polysemantic, dense)
Superposition $\sum_{\lambda} c_{\lambda} \lambda\rangle$	SAE expansion $\sum_{i \in S} a_i \mathbf{e}_i$
Coefficient $c_{\lambda} = \langle \lambda \psi \rangle$	SAE activation $a_i = \mathbf{e}_i \cdot r$
Orthonormal basis $\langle \lambda \lambda' \rangle = \delta_{\lambda \lambda'}$	SAE features approximately sparse but not orthogonal
Any orthonormal basis is valid	Dictionary $\{\mathbf{e}_i\}$ is over-complete ($D \gg d$)
Born rule: project onto eigenstate basis	Unembedding: project residual stream onto vocabulary

The key difference is orthogonality. In quantum mechanics, the monosemantic basis vectors are *orthonormal*: $\langle \lambda | \lambda' \rangle = \delta_{\lambda \lambda'}$. The expansion (2.46) is unique (it is a PCA-like decomposition in an orthonormal frame), and the coefficients $|c_{\lambda}|^2$ sum to exactly 1. In the transformer, the SAE dictionary vectors $\{\mathbf{e}_i\}$ are not orthogonal: the dictionary is over-complete ($D \gg d$), and the expansion (2.49) is not unique. Quantum mechanics uses a stricter, cleaner version of the same idea: polysemantic dense vectors expanded as superpositions of monosemantic orthogonal directions.

Dense vectors operated on by dense matrices. Both systems share a deeper structural similarity: the dynamics is entirely in terms of dense vectors and dense matrices.

In the transformer, the residual stream $r^{(\ell)} \in \mathbb{R}^d$ is updated at each layer by a learned weight matrix:

$$r^{(\ell+1)} = r^{(\ell)} + F_{\ell}(r^{(\ell)}), \quad (2.50)$$

where F_{ℓ} is parametrised by dense matrices $W_Q, W_K, W_V, W_O, W_{\text{ff}}$. The rules of grammar, semantics, and reasoning are not written explicitly; they are implicit in the weight matrices, discovered by gradient descent. The residual stream is an iterative algorithm written in the language of dense vectors and dense matrices, not in the language of logical rules.

In quantum mechanics, the hidden state $|\psi(t)\rangle \in \mathcal{H}$ is updated by the Hamiltonian:

$$|\psi(t + dt)\rangle = (I - iH dt) |\psi(t)\rangle, \quad (2.51)$$

where H is a dense matrix (the Hamiltonian operator). The rules of particle physics — scattering, decay, binding, confinement — are not written explicitly; they are implicit in the matrix elements of H . The Fock hidden state is an iterative algorithm written in the language of dense vectors (the Fock state) and dense matrices (the Hamiltonian), applied one Planck step at a time.

The two systems are not merely analogous. They are the same mathematical structure — iterative application of dense matrices to dense vectors — operating at vastly different scales. The transformer has $\sim 10^{11}$ parameters in its weight matrices and runs for $N \sim 10^2$ layers. The

Standard Model has 19 parameters in its Hamiltonian and runs for $\sim 10^{61}$ Planck steps on a Fock space of dimension $\gtrsim 2^{10^{186}}$. The quantum universe is incomparably larger, but not structurally different: both are polysemantic dense vectors, carrying a superposition of monosemantic basis states, rotated by dense matrices whose implicit rules generate all the richness we observe.

2.10 Making Fock space concrete: direct product, direct sum, and the full construction

Readers from machine learning may harbour a suspicion that the Fock space, the creation and annihilation operators, and the Dirac kets are abstract inventions—formal machinery that cannot be implemented as actual arrays and matrices. This suspicion is not unreasonable: most textbooks present the Fock space in Dirac notation from the start and never show you what it looks like as a column vector in memory.

This subsection dispels the suspicion. We construct the Fock space from scratch as concrete vectors and matrices, using two operations that have direct implementations in `numpy` or `PyTorch`:

Operation	QM name	ML name	Code
$\otimes$	Direct product (tensor product)	Outer product, then vectorise	<code>np.kron(a,b)</code>
$\oplus$	Direct sum	Concatenation with zero-padding	<code>np.block</code>

Once the construction is complete, the Dirac notation will be not a leap of faith but a shorthand for something you can type into a terminal.

Step 1: the direct product builds multi-index basis vectors

The simplest example: position and spin. Consider a single particle with two degrees of freedom: a position $x \in \{1, 2, \dots, 10\}$ (discretised for simplicity) and a spin $s \in \{\uparrow, \downarrow\}$. These live in two separate Hilbert spaces:

$$\mathcal{H}_x = \mathbb{C}^{10} \quad \text{with basis } \{|x\rangle\}_{x=1}^{10}, \quad \mathcal{H}_s = \mathbb{C}^2 \quad \text{with basis } \{|\uparrow\rangle, |\downarrow\rangle\}. \quad (2.52)$$

Each $|x\rangle$ is a 10-dimensional one-hot vector (a 1 in position x , zeros elsewhere). Each $|s\rangle$ is a 2-dimensional one-hot vector.

Their direct product $|x\rangle \otimes |s\rangle = |x, s\rangle$ is a basis vector of the joint space $\mathcal{H}_x \otimes \mathcal{H}_s$. We build it in three equivalent representations, each useful for different purposes.

Representation 1: the outer product (a matrix). The direct product of two column vectors $\mathbf{u} \in \mathbb{C}^m$ and $\mathbf{v} \in \mathbb{C}^n$ is their outer product: the $m \times n$ matrix

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u} \mathbf{v}^T = \begin{pmatrix} u_1 v_1 & u_1 v_2 & \cdots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \cdots & u_2 v_n \\ \vdots & & & \vdots \\ u_m v_1 & u_m v_2 & \cdots & u_m v_n \end{pmatrix}. \quad (2.53)$$

Each entry is the product of one entry from $\mathbf{u}$ and one entry from $\mathbf{v}$. In `numpy`: `np.outer(u, v)`.

For one-hot vectors, the outer product is a *one-hot matrix*: a matrix with a single 1 and all other entries 0. The entry at row x' , column s' is:

$$(|x\rangle \otimes |s\rangle)_{x', s'} = \delta_{x'x} \delta_{s's}. \quad (2.54)$$

For example, $|3\rangle \otimes |\uparrow\rangle = |3, \uparrow\rangle$ is the 10×2 matrix:

$$|3, \uparrow\rangle = |3\rangle \langle \uparrow| = \begin{array}{c|cc} & \uparrow & \downarrow \\ \hline 1 & 0 & 0 \\ 2 & 0 & 0 \\ 3 & 1 & 0 \\ 4 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 10 & 0 & 0 \end{array} \quad (2.55)$$

One entry is “hot”; the rest are cold. The row index tells you the position ($x = 3$); the column index tells you the spin ($s = \uparrow$). The two labels are packaged into one object.

Representation 2: vectorisation (the Kronecker product). A matrix is a two-index object. To use it as a vector in a larger linear algebra calculation, we *vectorise* it: stack the columns of the $m \times n$ matrix into a single column vector of length mn . This operation is called $\text{vec}(\cdot)$ in linear algebra and `.flatten()` or `.reshape(-1)` in code.

For the direct product, vectorisation produces the *Kronecker product*:

$$\text{vec}(\mathbf{u} \otimes \mathbf{v}) = \mathbf{u} \otimes_K \mathbf{v} = \begin{pmatrix} u_1 \mathbf{v} \\ u_2 \mathbf{v} \\ \vdots \\ u_m \mathbf{v} \end{pmatrix} = \begin{pmatrix} u_1 v_1 \\ u_1 v_2 \\ \vdots \\ u_1 v_n \\ u_2 v_1 \\ \vdots \\ u_m v_n \end{pmatrix} \in \mathbb{C}^{mn}. \quad (2.56)$$

Each entry of $\mathbf{u}$ is multiplied by the *entire* vector $\mathbf{v}$, and the results are stacked. In `numpy`: `np.kron(u, v)`.

For the position-spin example:

$$|3, \uparrow\rangle = \text{vec}(\text{one-hot matrix above}) = \underbrace{(0, 0, 1, 0, \dots, 0)}_{10 \text{ entries for } \uparrow}, \underbrace{(0, 0, 0, 0, \dots, 0)}_{10 \text{ entries for } \downarrow}^T \in \mathbb{C}^{20}. \quad (2.57)$$

This is a one-hot vector of length $10 \times 2 = 20$: a single 1 at position 3 in the $\uparrow$ -block.

Representation 3: the multi-way array (tensor). For more than two factors, the direct product is a multi-way array. Two particles, each with position and spin, give:

$$|x_1, s_1\rangle \otimes |x_2, s_2\rangle = |x_1, s_1, x_2, s_2\rangle, \quad (2.58)$$

which is a 4-index array of shape $10 \times 2 \times 10 \times 2$ —a tensor of order 4, one-hot at position (x_1, s_1, x_2, s_2) . Vectorising (flattening all indices into one) gives a one-hot vector of length $20^2 = 400$. In `numpy`: `np.kron(np.kron(e_x1, e_s1), np.kron(e_x2, e_s2))`.

For n particles: the tensor has order $2n$, shape $(10 \times 2)^n = 20^n$ after vectorisation. This exponential growth is real: it is the Hilbert space dimension, and it is why quantum simulation on classical hardware is exponentially expensive. The direct product is not a notational convenience—it is the source of the exponential complexity of quantum mechanics.

A general state is a weighted sum of one-hot products. A general state in $\mathcal{H}_x \otimes \mathcal{H}_s$ is:

$$|h\rangle = \sum_{x=1}^{10} \sum_{s=\uparrow, \downarrow} c(x, s) |x, s\rangle, \quad (2.59)$$

a superposition of all 20 one-hot basis vectors, with complex coefficients $c(x, s)$. In the matrix representation, $|h\rangle$ is a 10×2 matrix of complex numbers; in the vectorised representation, it is a dense vector in $\mathbb{C}^{20}$.

This is exactly the same as a 2D embedding table in a transformer: each token (x, s) has a one-hot representation of length 20 and a dense representation $c(x, s)$ obtained by matrix multiplication with the embedding table. The direct product builds the one-hot encoding for multi-index tokens.

Dense vectors work the same way. The basis vectors $|x\rangle$ and $|s\rangle$ need not be one-hot. If $|\phi\rangle$ is a dense 10-dimensional vector (a superposition in position) and $|\chi\rangle$ is a dense 2-dimensional vector (a superposition in spin), their Kronecker product $|\phi\rangle \otimes_K |\chi\rangle$ is the vectorised outer product:

$$(|\phi\rangle \otimes_K |\chi\rangle)_{(x,s)} = (|\phi\rangle)_x \cdot (|\chi\rangle)_s = \phi_x \chi_s. \quad (2.60)$$

A product of dense vectors is still a valid vector in the direct product space. We know from the subsection on basis agnosticism that the notation is basis-agnostic: whether the inputs are one-hot or dense, the Kronecker product and all the formalism built on it work identically.

Not every joint state is a product. A general state $|h\rangle = \sum_{x,s} c(x, s) |x, s\rangle$ does not necessarily factor as $|h\rangle = |\phi\rangle \otimes |\chi\rangle$. The factorisation holds if and only if the coefficient matrix $c(x, s)$ has rank 1: $c(x, s) = \phi_x \chi_s$ for some vectors ϕ, χ . When the rank exceeds 1, the state is *entangled*: it cannot be decomposed as a product of a position part and a spin part. In ML terms: a rank-1 matrix $c(x, s)$ is a separable embedding; a higher-rank matrix encodes correlations between position and spin that cannot be captured by separate embeddings.

Step 2: the direct sum is concatenation

The direct product gives us the n -particle Hilbert space for each fixed n . The Fock space must contain *all* particle numbers simultaneously: $n = 0, 1, 2, \dots$. This is where the direct sum enters.

The direct sum is concatenation with block structure. Given two vector spaces $\mathcal{H}_1 = \mathbb{C}^m$ and $\mathcal{H}_2 = \mathbb{C}^n$, their direct sum $\mathcal{H}_1 \oplus \mathcal{H}_2 = \mathbb{C}^{m+n}$ is the space of all vectors of length $m+n$, where the first m entries belong to $\mathcal{H}_1$ and the last n entries belong to $\mathcal{H}_2$:

$$|h\rangle = \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} \in \mathbb{C}^{m+n}, \quad \mathbf{h}_1 \in \mathbb{C}^m, \quad \mathbf{h}_2 \in \mathbb{C}^n. \quad (2.61)$$

In code: `np.concatenate([h1, h2])`. The direct sum is concatenation.

To embed a basis vector of $\mathcal{H}_1$ into the direct sum, we zero-pad: append n zeros at the bottom. To embed a basis vector of $\mathcal{H}_2$, we prepend m zeros at the top:

$$|x\rangle_{\oplus} = \underbrace{(0, \dots, 1, \dots, 0)}_{m \text{ entries}}, \underbrace{(0, \dots, 0)}_{n \text{ zeros}})^T \in \mathbb{C}^{m+n}, \quad (2.62)$$

$$|y\rangle_{\oplus} = \underbrace{(0, \dots, 0)}_{m \text{ zeros}}, \underbrace{(0, \dots, 1, \dots, 0)}_{n \text{ entries}})^T \in \mathbb{C}^{m+n}. \quad (2.63)$$

The two sets of basis vectors live in non-overlapping blocks and are therefore automatically orthogonal: $\langle x_{\oplus} | y_{\oplus} \rangle = 0$ for any x, y .

Concrete example. Let $\mathcal{H}_1 = \mathbb{C}^3$ with basis $|a\rangle, |b\rangle, |c\rangle$, and $\mathcal{H}_2 = \mathbb{C}^2$ with basis $|p\rangle, |q\rangle$. The

direct sum $\mathcal{H}_1 \oplus \mathcal{H}_2 = \mathbb{C}^5$ has five basis vectors:

$$|a\rangle_{\oplus} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, |b\rangle_{\oplus} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, |c\rangle_{\oplus} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, |p\rangle_{\oplus} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, |q\rangle_{\oplus} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \quad (2.64)$$

A general vector has amplitude in both blocks simultaneously:

$$|h\rangle = \alpha_a |a\rangle_{\oplus} + \alpha_b |b\rangle_{\oplus} + \alpha_c |c\rangle_{\oplus} + \beta_p |p\rangle_{\oplus} + \beta_q |q\rangle_{\oplus} = \begin{pmatrix} \alpha_a \\ \alpha_b \\ \alpha_c \\ \beta_p \\ \beta_q \end{pmatrix}. \quad (2.65)$$

The first three entries are the $\mathcal{H}_1$ -component; the last two are the $\mathcal{H}_2$ -component. The norm is $|\alpha_a|^2 + |\alpha_b|^2 + |\alpha_c|^2 + |\beta_p|^2 + |\beta_q|^2 = 1$.

What the direct sum means physically. In the Fock space context: $\mathcal{H}_1$ is the 0-particle sector, $\mathcal{H}_2$ is the 1-particle sector, $\mathcal{H}_3$ is the 2-particle sector, and so on. The direct sum $\mathcal{F} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \dots$ concatenates them all into one long vector. A general Fock state has amplitude in every sector simultaneously: the particle number is not fixed but is part of the superposition. Measuring the particle number operator is a Born-layer act that projects h_t onto one sector and trims away the rest.

Direct product vs. direct sum, compared.

	Direct product $\otimes$	Direct sum $\oplus$	Code
Operation	Outer product, then vectorise	Concatenation with zero-padding	<code>np.kron</code> vs. <code>np.concatenate</code>
Dimensions	Multiply: $m \times n$	Add: $m + n$	$\dim(\otimes) = mn$, $\dim(\oplus) = m + n$
Physical meaning	Combine degrees of freedom of one system (position $\times$ spin, or particle 1 $\times$ particle 2)	Combine sectors of different type (0-particle + 1-particle + 2-particle + $\dots$)	
Result	Multi-index basis vector (one configuration)	Superposition across sectors	

The Fock space combines both: direct products build the n -particle sectors; the direct sum assembles the sectors into one space.

Step 3: the Fock space is a direct sum of direct products

We now combine Steps 1 and 2 to build the full Fock space.

For a single-particle Hilbert space $\mathcal{H}_1$ of dimension d (e.g., $d = 20$ for our position-spin example), the n -particle Hilbert space is the n -fold direct product:

$$\mathcal{H}^{(n)} = \underbrace{\mathcal{H}_1 \otimes \mathcal{H}_1 \otimes \dots \otimes \mathcal{H}_1}_{n \text{ times}}, \quad \dim \mathcal{H}^{(n)} = d^n. \quad (2.66)$$

(For identical particles, we restrict to the symmetric or antisymmetric subspace; the construction is the same, with the dimension reduced. We discuss symmetrisation below.)

The Fock space is the direct sum (concatenation) over all particle numbers:

$$\mathcal{F} = \mathcal{H}^{(0)} \oplus \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)} \oplus \cdots = \bigoplus_{n=0}^{\infty} \mathcal{H}^{(n)}. \quad (2.67)$$

The 0-particle space $\mathcal{H}^{(0)} = \mathbb{C}$ is one-dimensional: its sole basis vector is the vacuum $|0\rangle$, a single complex number (the amplitude for “no particles”).

The Fock vector is a concatenation of vectors from each sector:

$$|h\rangle = \begin{pmatrix} c_0 \\ c_1(1) \\ c_1(2) \\ \vdots \\ c_1(d) \\ \hline c_2(1,1) \\ c_2(1,2) \\ \vdots \\ c_2(d,d) \\ \hline \vdots \end{pmatrix} \leftarrow \begin{array}{l} \mathcal{H}^{(0)} : 1 \text{ entry} \\ \mathcal{H}^{(1)} : d \text{ entries} \\ \mathcal{H}^{(2)} : d^2 \text{ entries} \\ \vdots \end{array} \quad (2.68)$$

Each block is zero-padded so that it occupies a non-overlapping segment of the infinite column vector. The vacuum $|0\rangle$ is the vector with $c_0 = 1$ and all other entries zero. A one-particle state $|k\rangle$ has $c_0 = 0$, a single nonzero entry $c_1(k) = 1$ in the $\mathcal{H}^{(1)}$ block, and zeros in all higher blocks.

A general Fock state is a superposition across all blocks:

$$|h\rangle = c_0 |0\rangle + \sum_{x,s} c_1(x,s) |x,s\rangle + \sum_{x_1,s_1,x_2,s_2} c_2(x_1,s_1,x_2,s_2) |x_1,s_1,x_2,s_2\rangle + \cdots \quad (2.69)$$

This vector has amplitude in the vacuum sector, amplitude in the one-particle sector, amplitude in the two-particle sector, and so on. The particle number is not fixed—it is part of the superposition.

Step 4: creation and annihilation as concrete matrices

The creation operator $a_k^\dagger$ and the annihilation operator a_k are matrices acting on the stacked Fock vector (2.68). Their action is to move amplitude between adjacent particle-number blocks.

Creation: moves from the n -particle block to the $(n+1)$ -particle block. The matrix $a_k^\dagger$ has nonzero entries only in the off-diagonal blocks connecting $\mathcal{H}^{(n)}$ to $\mathcal{H}^{(n+1)}$:

$$a_k^\dagger : \mathcal{H}^{(n)} \rightarrow \mathcal{H}^{(n+1)} \quad \text{for each } n. \quad (2.70)$$

Acting on a basis vector in the n -particle block, it produces a basis vector in the $(n+1)$ -particle block with one additional particle in mode k .

Annihilation: moves from the n -particle block to the $(n-1)$ -particle block.

$$a_k : \mathcal{H}^{(n)} \rightarrow \mathcal{H}^{(n-1)} \quad \text{for each } n \geq 1, \quad a_k |0\rangle = 0. \quad (2.71)$$

Acting on the vacuum gives zero: you cannot remove a particle from the empty state.

The normalization factor: why $\sqrt{n+1}$. The creation operator is defined to satisfy $[a_k, a_k^\dagger] = 1$. This commutation relation determines the normalization of $a^\dagger |n\rangle$. Here is the derivation:

$$\begin{aligned} \|a^\dagger |n\rangle\|^2 &= \langle n| a a^\dagger |n\rangle = \langle n| (a^\dagger a + [a, a^\dagger]) |n\rangle \\ &= \langle n| (a^\dagger a + 1) |n\rangle = \langle n| (\hat{n} + 1) |n\rangle = n + 1, \end{aligned} \quad (2.72)$$

where we used $[a, a^\dagger] = 1$ to rewrite $aa^\dagger = a^\dagger a + 1$, and $\hat{n} |n\rangle = n |n\rangle$. Since $\|a^\dagger |n\rangle\|^2 = n + 1$ and $a^\dagger |n\rangle$ must be proportional to $|n+1\rangle$ (the $(n+1)$ -particle state), we have:

$$\boxed{a^\dagger |n\rangle = \sqrt{n+1} |n+1\rangle, \quad a |n\rangle = \sqrt{n} |n-1\rangle.} \quad (2.73)$$

The second relation follows by taking the adjoint of the first (or by repeating the argument with $\|a |n\rangle\|^2 = \langle n| a^\dagger a |n\rangle = n$). The $\sqrt{n+1}$ and $\sqrt{n}$ factors are not conventions: they are consequences of the commutation relation.

The excited states from the vacuum. Applying $a^\dagger$ repeatedly:

$$|n\rangle = \frac{(a^\dagger)^n}{\sqrt{n!}} |0\rangle, \quad (2.74)$$

where the $\sqrt{n!}$ in the denominator accumulates from the $\sqrt{n+1}$ factors at each step: $a^\dagger |0\rangle = \sqrt{1} |1\rangle$, $a^\dagger |1\rangle = \sqrt{2} |2\rangle$, $\dots$, $(a^\dagger)^n |0\rangle = \sqrt{n!} |n\rangle$.

The block matrix structure. In the stacked representation, $a^\dagger$ and a are infinite-dimensional matrices with a characteristic block structure shifted by one block:

$$a^\dagger \sim \begin{pmatrix} 0 & 0 & 0 & \cdots \\ A^{(0 \rightarrow 1)} & 0 & 0 & \cdots \\ 0 & A^{(1 \rightarrow 2)} & 0 & \cdots \\ 0 & 0 & A^{(2 \rightarrow 3)} & \cdots \\ \vdots & & & \ddots \end{pmatrix}, \quad a \sim \begin{pmatrix} 0 & A^{(1 \rightarrow 0)} & 0 & \cdots \\ 0 & 0 & A^{(2 \rightarrow 1)} & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & & & \ddots \end{pmatrix}, \quad (2.75)$$

where $A^{(n \rightarrow n+1)}$ is a rectangular matrix of size $\dim \mathcal{H}^{(n+1)} \times \dim \mathcal{H}^{(n)}$ that maps an n -particle basis vector to the appropriate $(n+1)$ -particle basis vector, with the $\sqrt{n+1}$ factor. For a single mode (one harmonic oscillator), truncated to the first 5 levels, the matrices are those of Section 6:

$$a^\dagger = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \sqrt{1} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{4} & 0 \end{pmatrix}, \quad a = \begin{pmatrix} 0 & \sqrt{1} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{3} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{4} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (2.76)$$

These are actual matrices you can multiply. They are $a^\dagger$ and a . There is nothing more to creation and annihilation operators than these matrices.

The Hamiltonian $H = \sum_k \omega_k a_k^\dagger a_k$ is a product of these matrices: it is block-diagonal (the number operator $\hat{n} = a^\dagger a$ does not change the particle number), with each block giving the energy of the corresponding particle-number sector.

Step 5: for identical particles, symmetrise or antisymmetrise

In nature, identical particles are indistinguishable: the state $|k_1, k_2\rangle$ (particle 1 in mode k_1 , particle 2 in mode k_2) is physically the same as $|k_2, k_1\rangle$ (labels swapped). The Fock space for

identical particles restricts to the symmetric (bosons) or antisymmetric (fermions) subspace of each $\mathcal{H}^{(n)}$:

For **bosons**: $|k_1, k_2\rangle = +|k_2, k_1\rangle$ (symmetric under exchange). The two-particle basis vectors are $\frac{1}{\sqrt{2}}(|k_1\rangle \otimes |k_2\rangle + |k_2\rangle \otimes |k_1\rangle)$ for $k_1 \neq k_2$, and $|k\rangle \otimes |k\rangle$ for $k_1 = k_2$ (two particles in the same mode—allowed for bosons).

For **fermions**: $|k_1, k_2\rangle = -|k_2, k_1\rangle$ (antisymmetric under exchange). Setting $k_1 = k_2$: $|k, k\rangle = -|k, k\rangle = 0$ —two fermions cannot occupy the same mode. This is the Pauli exclusion principle, derived from the antisymmetry, not postulated.

The symmetrisation reduces the dimension of $\mathcal{H}^{(n)}$ from d^n to $\binom{d+n-1}{n}$ (bosons) or $\binom{d}{n}$ (fermions). The creation operators automatically enforce the correct symmetry: $a_k^\dagger$ adds a particle in a way that maintains (anti)symmetry, with the appropriate sign for fermions.

Step 6: the construction is one instantiation

The stacked, zero-padded representation above is one concrete instantiation of the Fock space—the one that makes the direct-sum and direct-product structure visible. But we know from the subsection on basis agnosticism that we can work in *any* orthonormal basis.

In principle, we could follow the transformer paradigm: embed each Fock basis vector as a dense vector in some $\mathbb{R}^{d_{\text{emb}}}$ ($d_{\text{emb}} \sim 10^4$, say), and represent the creation and annihilation operators as dense matrices in this embedding space. The stacked one-hot construction would be one particular basis (the “token index”); the dense embedding would be another. The physics—the inner products, the Born probabilities, the Schrödinger evolution—would be identical in both, because it depends only on orthonormality and completeness.

The explicit construction of Steps 1–5 serves a different purpose: it proves that the Fock space, the creation and annihilation operators, and the Dirac notation are not empty formalism. They can be written down as actual arrays and actual matrices, with entries you can compute and verify. The one-hot stacked representation is the proof of existence. The Dirac notation is the language we use once the existence is established.

Step 7: from concrete vectors to Dirac kets

With the construction complete, we can now use Dirac notation with confidence. The symbol $|n_{\mathbf{k}_1\lambda_1}, n_{\mathbf{k}_2\lambda_2}, \dots\rangle$ is a shorthand for a specific stacked, zero-padded, one-hot vector in the infinite-dimensional Fock space (2.68). The symbol $a_{\mathbf{k}\lambda}^\dagger$ is a shorthand for the block-shifted matrix (2.75). The Schrödinger equation $\dot{h}_t = -iHh_t$ is a matrix-vector multiplication on the stacked vector. The Born rule $P(\lambda) = \|P_\lambda h_t\|^2$ is the squared Euclidean norm of a subvector.

Dirac notation	Concrete implementation
$ n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$	One-hot vector in a stacked column
$a_k^\dagger$	Block-shifted matrix with $\sqrt{n+1}$ entries
$\dot{h}_t = -iHh_t$	Matrix-vector multiply on the stacked vector
$P(\lambda) = \ P_\lambda h_t\ ^2$	Squared norm of a subvector (projection)
$\mathcal{H}^{(n_1)} \otimes \mathcal{H}^{(n_2)}$	Kronecker product: <code>np.kron(h1, h2)</code>
$\mathcal{H}^{(0)} \oplus \mathcal{H}^{(1)} \oplus \dots$	Concatenation: <code>np.concatenate([h0, h1, ...])</code>

We will never write these stacked vectors or block matrices again. The Dirac notation— $|n\rangle$, $a^\dagger$, a , $\langle n|h_t\rangle$ —does everything the explicit construction does, in a form that is simultaneously

more compact, more general (basis-agnostic), and more readable. But behind every Dirac ket is a concrete vector; behind every operator is a concrete matrix; and behind the Fock space is the direct sum of direct products (concatenation of Kronecker products) constructed in this subsection. The notation is not a story. It is a compression of a construction you have now seen—and can implement.

3 Quantum Mechanics as an RNN: The Core Model

3.1 The Fock-Born architecture

Linear RNNs in deep learning. A recurrent neural network (RNN) maintains a hidden state $h_t \in \mathbb{R}^d$ that evolves over a sequence of time steps $t = 0, 1, 2, \dots$ by a fixed rule:

$$h_{t+1} = f(Wh_t + Ux_t + b), \quad (3.1)$$

where $W \in \mathbb{R}^{d \times d}$ is the recurrent weight matrix, x_t is the input at step t , U and b are further parameters, and f is a nonlinearity. The hidden state is a dense vector: it encodes, in a distributed representation, everything the network has learned from the sequence so far. Between output steps, the hidden state is internal—it has no direct interpretation in the vocabulary of observable tokens.

A *linear* RNN drops the nonlinearity and the input:

$$h_{t+1} = Wh_t \implies h_t = W^t h_0. \quad (3.2)$$

The hidden state at time t is the matrix power W^t applied to the initial state. This is extremely simple—and extremely relevant, because recent architectures (S4, Mamba, RWKV, Griffin, and other state-space models) have shown that linear or near-linear recurrences, when properly parameterised, achieve competitive performance with transformers on many sequence-modelling tasks.

The quantum Schrödinger equation is a continuous-time linear RNN with a specific constraint on the weight matrix.

The Fock layer is the game engine. It is one equation:

$h_t = e^{-iHt} h_0.$

(3.3)

A unit vector h_t in a complex Hilbert space $\mathcal{H}$, evolving under a fixed Hermitian operator H . Comparing with the linear RNN (3.2):

	Linear RNN	Schrödinger RNN
Hidden state	$h_t \in \mathbb{R}^d$	$h_t \in \mathbb{C}^d$ (complex)
Discrete step	$h_{t+1} = Wh_t$	$h_{t+\epsilon} = (I - iH\epsilon)h_t$
Accumulated evolution	$W^t h_0$	$e^{-iHt} h_0$
Weight matrix	W (arbitrary)	$e^{-iH\epsilon}$ (unitary)
Norm of h_t	Can grow or shrink	$\ h_t\ = 1$ always
Eigenvalue of W	$ \lambda_k $ (magnitude, can vary)	$ e^{-iE_k\epsilon} = 1$ (unit circle)

The three differences from a standard linear RNN are:

(i) *Complex-valued.* The hidden state and weight matrix are complex. This is not optional: the factor i in $W = -iH$ converts real eigenvalues of H into purely imaginary eigenvalues of W , which generate rotations ($|e^{-iE_k t}| = 1$) rather than exponential growth or decay ($|e^{\lambda_k t}| \neq 1$). If H is real and i is absent, eigenvalues grow and the norm blows up. The complex structure of $\mathcal{H}$ is what makes the engine conserve probability.

(ii) *Unitary.* The evolution e^{-iHt} is a unitary matrix: $\|h_t\| = \|h_0\| = 1$ for all t . The hidden state rotates on the unit sphere in Hilbert space; it never grows or shrinks. This guarantees that the Born probabilities sum to 1 at every instant without any explicit normalisation. A general

linear RNN can have eigenvalues inside or outside the unit circle, causing the hidden state to decay or explode. The Schrödinger RNN sits exactly on the unit circle—every eigenvalue has $|e^{-iE_k\epsilon}| = 1$.

(iii) *No input.* The evolution $h_t = e^{-iHt}h_0$ has no external input sequence x_t . The engine is autonomous: once the initial state h_0 and the Hamiltonian H are specified, the evolution is determined for all time. (The Born layer provides the only interaction with the outside: the observer submits an observable and receives an outcome. This is more like a read/write interface than an input sequence.)

The Fock layer is a vector rotating. The rotation is continuous, deterministic, unitary, and linear. The vector never changes length. The operator never changes. That is all.

The Born layer is the interactive interface between the game engine and the observer. It is three lines:

$$(1) \text{ Born probability: } p(\lambda) = \sum_{k=1}^{d_\lambda} |\langle \lambda, k | \psi(t) \rangle|^2, \quad (3.4)$$

$$(2) \text{ Born sampling: } \lambda \sim p(\lambda), \quad (3.5)$$

$$(3) \text{ Bohr trimming: } |\psi(t^+)\rangle = \frac{P_\lambda |\psi(t)\rangle}{\sqrt{p(\lambda)}}. \quad (3.6)$$

The observer chooses an observable $\hat{O} = \sum_\lambda \lambda P_\lambda$, with eigenvalues $\{\lambda\}$ and eigensubspace projectors $\{P_\lambda\}$. Step (1) computes the probability of each eigenvalue by projecting h_t onto the corresponding eigensubspace and taking the squared length. Step (2) draws a sample from this distribution: the observer receives one classical number λ . Step (3) trims the hidden state: everything outside the eigensubspace of the observed λ is discarded, everything inside survives with its original amplitudes and phases, and the result is renormalised to unit length. The Schrödinger rotation then resumes from $|\psi(t^+)\rangle$.

A standard linear RNN has a fixed output layer: one unembedding matrix, one vocabulary, applied at every time step. The Born layer differs in two essential ways: it is *interactive* (the observer chooses the observable at measurement time) and *multi-lingual* (different observables define different vocabularies, different query vectors, and different probability distributions—all read from the same hidden state). We develop the Born layer in full detail in the next subsection.

One line for the engine. Three lines for the interface. This is the complete mathematical specification of non-relativistic quantum mechanics. Everything else in this book—Hamiltonians, Fock spaces, path integrals, Feynman diagrams, the Standard Model—is either a choice of H (specifying the engine) or a calculational tool for computing properties of e^{-iHt} (servicing the engine). The architecture itself is these four lines.

When the Born distribution is a delta. If the Born distribution $p(\lambda)$ is concentrated on a single eigenvalue— $p(\lambda_0) \approx 1$ and $p(\lambda) \approx 0$ for $\lambda \neq \lambda_0$ —then all three steps of the Born layer become trivial: the probability is near-certain, the sample is deterministic, and the trimming barely alters the hidden state. This is the classical regime. The Born interface still operates—an eigenvalue is still rendered, a sample is still drawn, the hidden state is still trimmed—but the outcome is a foregone conclusion. The distribution has already decided. The interface is inert.

A macroscopic observable $\hat{\Phi} = N^{-1} \sum_k \hat{\phi}_k$ acting on $N \sim 10^{23}$ degrees of freedom is a delta-distribution observable: its eigenvalue degeneracy is $\sim e^S$, its variance is $O(1/N)$, and its Born output is effectively deterministic. Classical physics is the regime where the Born interface is trivially passable—not because the Fock layer has changed its nature, but because the observer’s

chosen observable has a Born distribution so concentrated that sampling and trimming do nothing.

Warning: do not add drama to the Fock layer. Look at equation (3.3) again. There is a vector. It rotates. Nothing else happens.

No particles are emitted. No paths are taken. No fields fluctuate. No vacuum teems. No barriers are tunnelled through. No histories are summed over. These are all narrations borrowed from the Born layer—especially from observables whose Born distributions are near-delta, where the eigenvalues look so classical that they seem to be properties of the Fock layer itself. They are not. They are Born-layer renderings: outputs of the interface, not states of the engine.

Narration	Layer it belongs to	What actually happens
“The atom emits a photon”	Born (detection event)	h_t rotates continuously
“The electron tunnels through the barrier”	Born (position measured on both sides)	h_t rotates in Hilbert space
“Virtual particles pop in and out”	Born (basis-decomposition coefficients)	Vacuum $ \Omega\rangle$ sits still
“The field fluctuates”	Born (variance of field observable)	h_t is one definite vector
“The particle takes every path”	Born (path integral is a matrix computation)	h_t is one vector, not many paths

The temptation is strongest for observables with delta-like Born distributions, because their outputs are deterministic and repeatable—they look like facts about the system rather than outcomes of a measurement act. The position of a planet, the reading of a voltmeter, the value of a classical field—these all feel like properties of the world, not renderings through an interface. But in the Fock-Born architecture, they are renderings. They are eigenvalues of macroscopic observables whose Born distributions happen to be near-delta. The fact that sampling is trivial does not change the layer they belong to.

The diagnostic. Whenever a physical intuition narrates events inside the Fock layer—anything happening other than a vector rotating—apply this test: *is an observer submitting an observable and receiving an eigenvalue?* If not, the narration is smuggling Born-layer language into the wrong layer.

The Fock layer has no events. It has one vector, rotating under one Hamiltonian. All events belong to the Born layer.

3.2 The Fock layer: a single vector, rotating

From compound interest to quantum evolution. The Schrödinger equation is the matrix generalisation of compound interest.

Start with a scalar: a bank balance x_t growing at continuous rate r . Over a short interval dt :

$$x_{t+dt} = (1 + r dt) x_t. \quad (3.7)$$

Apply N times over $[0, T]$ with $\Delta t = T/N$:

$$x_T = (1 + r \Delta t)^N x_0 \xrightarrow{N \rightarrow \infty} e^{rT} x_0. \quad (3.8)$$

Now let the state be a vector $h_t \in \mathbb{C}^n$ and the rate a matrix $W \in \mathbb{C}^{n \times n}$:

$$h_{t+dt} = (I + W dt) h_t \implies h_T = e^{WT} h_0, \quad (3.9)$$

where the *matrix exponential* is defined by the same power series as the scalar one: $e^{WT} = \sum_{k=0}^{\infty} (WT)^k / k!$.

The Schrödinger equation is exactly this with $W = -iH$:

$$\dot{h}_t = -iH h_t \implies \boxed{h_t = e^{-iHt} h_0.} \quad (3.10)$$

Compound interest	Matrix ODE	Schrödinger
Scalar $x_t \in \mathbb{R}$	Vector $h_t \in \mathbb{C}^n$	Ket $ \psi(t)\rangle \in \mathcal{H}$
Rate $r \in \mathbb{R}$	Matrix W	Operator $-iH$
Solution $e^{rT} x_0$	Solution $e^{WT} h_0$	Solution $e^{-iHT} \psi(0)\rangle$
Growth: $ e^{rT} > 1$	Depends on eigenvalues	Rotation: $ e^{-iE_k t} = 1$

The i converts growth into rotation. A real symmetric H without the i would generate e^{-Ht} : an exponentially growing or decaying transformation. The $i = \sqrt{-1}$ converts each real eigenvalue E_k of H into a purely imaginary eigenvalue $-iE_k$ of $W = -iH$, and $e^{-iE_k t}$ has unit modulus for all t . Growth becomes rotation. The norm $\|h_t\|$ is exactly preserved, guaranteeing that the Born probabilities sum to 1 at every instant. The complex structure of $\mathcal{H}$ is not a convention; it is what makes the engine conserve probability.

The hidden state is basis-agnostic. The vector h_t does not live in any particular basis. It is simultaneously a superposition in every orthonormal basis of $\mathcal{H}$:

$$h_t = \sum_n \langle E_n | \psi(t) \rangle |E_n\rangle = \int dx \langle x | \psi(t) \rangle |x\rangle = \int \frac{d^3k}{(2\pi)^3} \langle \mathbf{k} | \psi(t) \rangle |\mathbf{k}\rangle = \dots \quad (3.11)$$

These are not different states. They are the same vector, decomposed in different bases—energy, position, momentum, or any other. The coefficients $\langle E_n | \psi(t) \rangle$, $\langle x | \psi(t) \rangle$, $\langle \mathbf{k} | \psi(t) \rangle$ are the same vector's components in different coordinate systems.

This is the hidden state's defining property: it is not committed to any vocabulary until the Born layer is invoked. The game engine runs without knowing what language the observer will speak. The vector h_t carries all possible decompositions simultaneously, and no decomposition is privileged inside the Fock layer. Only the Born layer, by choosing an observable, selects a basis in which to read the vector off.

3.3 The Hamiltonian: the motor of the game engine

The hidden state h_t is the car. The Hamiltonian H is the engine. It is the single most important object in the theory: it determines the rotation, the spectrum, the dynamics, the energy levels, the spectral lines, the scattering amplitudes—everything. Once H is specified, the Fock layer is fully determined.

The entire content of a quantum theory can be stated in two questions: *What is the Hilbert space $\mathcal{H}$?* and *What is the Hamiltonian H ?* The rest—evolution, spectra, transition rates, thermodynamics—follows from $h_t = e^{-iHt} h_0$ and the Born layer.

Common forms. Every Hamiltonian encountered in this book falls into one of two classes:

Kinetic plus potential. For a single particle in a potential:

$$H = \frac{\hat{p}^2}{2m} + V(\hat{x}). \quad (3.12)$$

The Hilbert space is $\mathcal{H} = L^2(\mathbb{R}^3)$, the space of square-integrable functions on $\mathbb{R}^3$. The basis vectors are position eigenstates $|\mathbf{x}\rangle$ or momentum eigenstates $|\mathbf{k}\rangle$, related by the Fourier transform. The potential $V(\hat{x})$ encodes the physics: harmonic oscillator ($V = \frac{1}{2}m\omega^2 x^2$), hydrogen atom ($V = -e^2/r$), particle in a box ($V = 0$ inside, ∞ outside).

Polynomials of creation and annihilation operators. For systems with variable particle number:

$$H = \sum_k \omega_k a_k^\dagger a_k + H_{\text{int}}(a^\dagger, a). \quad (3.13)$$

The Hilbert space is the Fock space $\mathcal{F} = \bigoplus_{n=0}^{\infty} \mathcal{H}^{(n)}$, the direct sum over all particle-number sectors. The operator $a_k^\dagger$ creates a particle in mode k ; a_k annihilates one. The first term counts the energy of non-interacting modes; H_{int} adds interactions (scattering, pair creation, decay). This is the language of condensed matter physics (Hubbard, Heisenberg, BCS, Bose–Hubbard) and of quantum field theory (the Standard Model). In condensed matter, one writes H directly and never passes through a Lagrangian.

The eigen structure: rotation rates of the engine. H is Hermitian, so the spectral theorem gives:

$$H = \sum_n E_n |E_n\rangle \langle E_n|, \quad H |E_n\rangle = E_n |E_n\rangle. \quad (3.14)$$

The eigenvalues $\{E_n\}$ are the *energy levels*—real numbers, because H is Hermitian. The eigenstates $\{|E_n\rangle\}$ are the directions in Hilbert space that do not mix under evolution: each simply acquires a phase,

$$e^{-iHt} |E_n\rangle = e^{-iE_n t} |E_n\rangle. \quad (3.15)$$

The eigenvalue E_n is the rotation rate of eigenstate $|E_n\rangle$ in the complex plane. A general state $h_0 = \sum_n c_n |E_n\rangle$ evolves as

$$h_t = \sum_n c_n e^{-iE_n t} |E_n\rangle : \quad (3.16)$$

each mode rotates independently at its own frequency E_n . The dynamics of the Fock layer is completely determined by the eigenvalues and eigenstates of H —nothing else.

Discrete spectrum and the word “quantum.” For a free particle, the eigenvalues form a continuous band: $E = \mathbf{k}^2/2m$ for any $\mathbf{k} \in \mathbb{R}^3$. But for a bound system—a particle confined by a potential well, an electron in a hydrogen atom, a mass on a spring—the boundary conditions force the eigenvalues to be discrete:

$$E_0 < E_1 < E_2 < \dots \quad (3.17)$$

Harmonic oscillator: $E_n = \omega(n + \frac{1}{2})$, equally spaced. Hydrogen atom: $E_n = -13.6 \text{ eV}/n^2$, converging to zero. Between consecutive levels, no energy is allowed. The energy is *quantised*—it comes in discrete amounts. This is the *quantum* in quantum mechanics.

Discrete eigenvalues mean discrete rotation rates. The engine can only oscillate at the specific frequencies $E_0, E_1, E_2, \dots$, not at any intermediate frequency. Atoms emit light only at specific frequencies because the photon frequency is the difference of two rotation rates: $\nu = (E_m - E_n)/h$. Atomic spectral lines are the fingerprint of the discrete rotation rates of the Hamiltonian motor.

The word “spectral.” The *spectrum* of H is the set of its eigenvalues $\{E_n\}$. This is the same word Newton used for the band of colours from a prism (Latin *spectrum*: ghost, apparition), and the same word Kirchhoff used for the discrete emission lines of heated elements. Hilbert borrowed the word for the eigenvalues of an operator, intending it as a metaphor. Quantum mechanics revealed the metaphor to be an identity: the physical spectrum (Kirchhoff’s spectral lines) IS

the mathematical spectrum (Hilbert’s eigenvalues). The colours emitted by hydrogen are the eigenvalues of the hydrogen Hamiltonian. The spectral decomposition $H = \sum_n E_n |E_n\rangle \langle E_n|$ is literally a decomposition into spectral lines.

The Hamiltonian is the Fock layer’s only free parameter. Once the Hilbert space $\mathcal{H}$ and the Hamiltonian H are specified, the game engine is fully determined. The evolution e^{-iHt} is fixed. The spectrum $\{E_n\}$ is fixed. Every property of the Fock layer follows. The entire programme of quantum mechanics, quantum field theory, and the Standard Model is the programme of writing down the correct H —the correct motor—and then computing its consequences. The Lagrangian, the path integral, Feynman diagrams, and renormalisation are all tools for this computation. The Hamiltonian is the object they serve.

3.4 The Born layer: a multi-lingual interactive output layer

A standard RNN has a fixed output layer: one unembedding matrix W_U , one vocabulary $\{w_1, \dots, w_V\}$, applied identically at every time step. The output logit for token w is the inner product $\ell_w = \langle W_U[w], h_t \rangle$ between the hidden state and the w -th row of the unembedding matrix. The probability is $p(w) = \text{softmax}(\ell)_w$.

The Born layer is a radically more powerful version of this output layer. It differs from the standard RNN output in three ways: it is *interactive* (the observer chooses the unembedding matrix at measurement time), it is *multi-lingual* (different choices define different vocabularies), and it handles *degeneracy* (multiple query vectors can map to the same output token).

The observable defines three things simultaneously

When the observer chooses an observable $\hat{O} = \sum_\lambda \lambda P_\lambda$, this single choice defines:

1. **The vocabulary:** the set of eigenvalues $\{\lambda_1, \lambda_2, \dots\}$. These are the classical words the Born layer can output—the possible readings on the observer’s dial. Each eigenvalue is a real number: a position, a momentum, an energy, a spin projection, a particle count.
2. **The query vectors:** the eigenvectors $\{|\lambda, k\rangle\}_{k=1}^{d_\lambda}$ for each eigenvalue λ . These are the directions in Hilbert space that the Born layer projects onto—the “keys” against which the hidden state is compared. The inner product $\langle \lambda, k | h_t \rangle$ is the attention score between query vector $|\lambda, k\rangle$ and the hidden state h_t .
3. **The degeneracy:** the number d_λ of query vectors per eigenvalue λ . When $d_\lambda > 1$, multiple orthogonal directions in Hilbert space produce the same classical output λ . The Born probability sums over all of them.

The three are packaged together in one Hermitian operator. Choosing the observable is choosing the read head.

The Born probability as an aggregated inner product. The Born probability for eigenvalue λ is:

$$p(\lambda) = \sum_{k=1}^{d_\lambda} |\langle \lambda, k | h_t \rangle|^2 = \sum_{k=1}^{d_\lambda} |\text{attention score with } k\text{-th query}|^2. \quad (3.18)$$

Each term $|\langle \lambda, k | h_t \rangle|^2$ is the squared inner product of h_t with one query vector—the “relevance” of h_t to that direction. The sum over k aggregates all query vectors that produce the same output token λ .

This is an *aggregated squared softmax*: the inner products $\langle \lambda, k | h_t \rangle$ play the role of logits, the

squared modulus plays the role of exponentiation, and the sum over k aggregates all queries that map to the same token. The crucial differences from a standard softmax are:

Standard (RNN/transformer)	softmax	Born rule (quantum)
$p(w) = \frac{e^{\ell_w}}{\sum_{w'} e^{\ell_{w'}}$		$p(\lambda) = \sum_k \langle \lambda, k h_t \rangle ^2$
One query vector per token		d_λ query vectors per eigenvalue (degeneracy)
Requires explicit partition function Z		Automatically normalised ($\sum_\lambda p(\lambda) = 1$)
Unembedding matrix is fixed		Observable (read head) chosen at measurement time

Automatic normalisation. The normalisation $\sum_\lambda p(\lambda) = 1$ is not computed; it is guaranteed by the architecture:

$$\sum_\lambda p(\lambda) = \sum_\lambda \sum_k |\langle \lambda, k | h_t \rangle|^2 = \langle h_t | \underbrace{\left(\sum_{\lambda, k} |\lambda, k\rangle \langle \lambda, k| \right)}_{=1} | h_t \rangle = \|h_t\|^2 = 1. \quad (3.19)$$

The completeness of the eigenbasis (which is guaranteed because $\hat{\mathcal{O}}$ is Hermitian) and the unitarity of the Schrödinger evolution (which preserves $\|h_t\| = 1$) together ensure that the Born probabilities sum to 1 at every instant. No partition function Z needs to be computed. The Born layer is the natural output layer for a unitary engine.

Degeneracy: the key to classical physics

The number of query vectors per eigenvalue—the degeneracy d_λ —controls the character of the Born output.

Non-degenerate ($d_\lambda = 1$ for all λ). Each eigenvalue has exactly one query vector. There is no aggregation. The Born probability is a single squared inner product: $p(\lambda) = |\langle \lambda | h_t \rangle|^2$. The interference between different λ 's is fully visible. A Born sample is genuinely random. The Bohr trimming resets h_t to a single eigenvector: maximal collapse. This is the quantum regime: full randomness, full interference, full information extraction.

Highly degenerate ($d_\lambda \gg 1$). Each eigenvalue is shared by many query vectors. The Born probability sums over all of them: $p(\lambda) = \sum_{k=1}^{d_\lambda} |\langle \lambda, k | h_t \rangle|^2$. The sum washes out the fine-grained information: the observer cannot distinguish which of the d_λ query vectors was responsible. Cross terms between different k 's vanish by orthogonality—no interference within an eigensubspace. The Born output is coarser, less informative, less random. When the degeneracy is enormous ($d_\lambda \sim e^S$, thermodynamic), the Born distribution concentrates into a near-delta spike, and the output is effectively deterministic. This is the classical regime.

Maximally degenerate ($d_\lambda = \dim \mathcal{H}$, one eigenvalue). The identity observable $\hat{\mathcal{O}} = \mathbf{1}$: every vector is a query vector for the single eigenvalue $\lambda = 1$. The Born probability is $p(1) = \|h_t\|^2 = 1$. No information is extracted, no trimming occurs. The measurement is vacuous.

The degeneracy d_λ is the single parameter that interpolates between the fully quantum and the fully classical. A later section develops this into the full theory of degenerate observables.

The observer defines the language

Different observables define different languages. Each language has its own vocabulary, its own query vectors, and its own degeneracy structure:

Observable	Vocabulary	Query vectors	d_λ	Character
Position $\hat{x}$	$\{x \in \mathbb{R}\}$	$ x\rangle$ (delta functions)	1	Quantum
Momentum $\hat{p}$	$\{p \in \mathbb{R}\}$	$ p\rangle$ (plane waves)	1	Quantum
Energy H	$\{E_n\}$	$ E_n\rangle$	$1-g_n$	Varies
Spin- z $\hat{S}_z$	$\{+\hbar/2, -\hbar/2\}$	$ \uparrow\rangle, \downarrow\rangle$	1	Quantum
Total particle number $\hat{N}$	$\{0, 1, 2, \dots\}$	All Fock states with n particles	Huge	Classical
Macroscopic field $\hat{\Phi}$	$\{\Phi_0\}$ (near-definite)	$\sim e^S$ microstates	$\sim e^S$	Classical
Identity 1	$\{1\}$	Every vector	$\dim \mathcal{H}$	Trivial

The hidden state h_t is the same object in every language. Only the Born-layer interface changes. This is the basis agnosticism of the Fock layer, expressed as the multi-lingual capability of the Born layer.

In ML terms. Imagine a transformer whose unembedding layer can be swapped at inference time. One unembedding matrix projects the residual stream onto an English vocabulary; another projects onto a French vocabulary; a third projects onto a Python-code vocabulary. The residual stream is the same vector; the choice of unembedding determines which language the output is in.

The Born layer goes further: the ML analogue would be an unembedding matrix where multiple rows (query vectors) are labelled by the same output token, so the probability of that token is the sum of the squared scores across all its query vectors. This aggregation is what “degeneracy” means, and it is how the Born layer washes out fine-grained information to produce classical-looking outputs from a quantum hidden state.

The three-step protocol, elaborated

The Born layer’s three operations (3.4)–(3.6) deserve a closer look in the vocabulary/query-vector language.

Step (1): Born probability (the readout). The observer declares an observable $\hat{O}$. This defines a vocabulary $\{\lambda\}$ and, for each λ , a set of d_λ query vectors $\{|\lambda, k\rangle\}$. The Born layer computes the inner product of h_t with every query vector, squares the moduli, and sums within each eigenvalue: $p(\lambda) = \sum_k |\langle \lambda, k | h_t \rangle|^2$. The result is a classical probability distribution over the vocabulary. This is the rendering: the game engine’s internal state is converted into a classical probability distribution over the observer’s chosen words.

Step (2): Born sampling (the output token). Nature draws a sample λ from $p(\lambda)$. The observer receives one classical number—a definite, readable value. This is the moment where the classical world appears: a single eigenvalue, recorded by the observer.

Step (3): Bohr trimming (the feedback to the engine). The hidden state is updated: project h_t onto the eigensubspace of the sampled λ , discard everything outside, and renormalise to unit length:

$$h_{t+} = \frac{P_\lambda h_t}{\|P_\lambda h_t\|} = \frac{\sum_{k=1}^{d_\lambda} |\lambda, k\rangle \langle \lambda, k | h_t \rangle}{\sqrt{\sum_{k=1}^{d_\lambda} |\langle \lambda, k | h_t \rangle|^2}}. \quad (3.20)$$

The post-measurement state lives in the eigensubspace of λ : it is a superposition of the d_λ query

vectors, with their original amplitudes and phases preserved (only the overall normalisation changes). If $d_\lambda = 1$ (non-degenerate), the state collapses to a single eigenvector: $h_{t+} = |\lambda\rangle$. If d_λ is large (degenerate), a rich superposition survives inside the eigensubspace: the trimming removes little, and the engine continues from a state that is only slightly altered.

In ML terms: Bohr trimming is like *hard top-1 decoding* followed by *re-embedding*. The output token λ is selected; the hidden state is reset to the subspace corresponding to that token (the “embedding” of the observed outcome); and the engine resumes. For a non-degenerate observable, the re-embedding is a hard lookup: $h_{t+} = |\lambda\rangle$, a single vector determined by the output token. For a degenerate observable, the re-embedding preserves the internal structure within the eigensubspace—a soft re-embedding.

The three steps fire together, as a single interaction between the observer and the engine.

Classical variables as a tokeniser

The eigenvalues $\{\lambda\}$ of the observer’s chosen observable are the classical vocabulary—the set of outcomes that the Born layer can render. This vocabulary is declared in advance, by the observer, exactly as a tokeniser declares the vocabulary before a language model runs. The Schrödinger equation knows nothing about this choice: it evolves h_t regardless of which observable the observer will eventually submit.

The classical outcome λ exists only when the Born layer is invoked. Between measurements, there is no classical variable. There is only h_t , rotating under the Hamiltonian. The question “what is the position of the electron between measurements?” is like asking “what was GPT saying at layer 40?” Layer 40 is an intermediate hidden state; it has no interpretation in the vocabulary of classical tokens. The question is not unanswerable; it is *malformed*: it asks for a Born-layer quantity (a classical eigenvalue) at a moment when the Born layer has not been invoked.

Between measurements, the screen is dark. When the game engine is running but no frame has been rendered, there is no image on the screen. The game world evolves internally, in its hidden state, but the classical rendering—the pixels, the colours, the shapes—is simply absent.

The classical world is the screen. Measurement is the rendering. Between renderings, the screen is dark. The hidden state continues to evolve—the engine runs—but the classical world is not there. This is not a limitation of our knowledge; it is a structural feature of the architecture. The classical world is not a substance or a container or a background; it is a sequence of rendered frames, produced on demand by the Born layer when the observer submits an observable.

Most of the famous paradoxes of quantum mechanics—Schrödinger’s cat, the EPR paradox, the measurement problem—arise from the assumption that the classical world exists continuously, even between measurements. In the Fock-Born architecture, this assumption is rejected at the level of the axioms. The cat’s alive-or-dead status is a classical variable that exists only when the Born layer is invoked with the alive/dead observable. The paradoxes dissolve not because a clever argument resolves them, but because the premise that generates them—the continuous existence of the classical world—is not part of the architecture.

3.5 The observer as a primitive

The observer is not part of the hidden state. The observer is the entity that submits an observable to the Born layer and receives a classical outcome. Attempting to model the observer as a subsystem of h_t —placing the observer inside the Hilbert space—is the same category error as modelling the GPT user as a vector in the residual stream. The user is outside the model; the

observer is outside the hidden state.

The Copenhagen interpretation as an axiomatic system. The Fock-Born architecture is not one interpretation of quantum mechanics among many. It is the axiomatic foundation of the Copenhagen interpretation, and the Copenhagen interpretation is not an interpretation in the philosophical sense. It is an *axiomatic system*, exactly as Euclidean geometry is.

In Euclidean geometry, the primitive objects—points, lines, planes—are intentionally left undefined. The axioms specify their relationships without defining what they are. A “point” is not explained by the geometry; it is a primitive that the geometry takes as given.

The Fock-Born architecture has the same structure. The *observer* is a primitive: intentionally left undefined, exactly as the point is left undefined in Euclidean geometry. The two axioms—the Schrödinger equation specifying hidden-state dynamics, and the Born protocol specifying the measurement interface—define the relationships between the observer and the hidden state without defining what an observer is. Asking “but what is the observer, really?” is like asking “but what is a point, really?” in Euclidean geometry: it is a question the axiomatic system deliberately does not answer, because the power of the system lies in the axioms, not in the definition of the primitives.

Deriving the output layer from the hidden layer is a category error. The Schrödinger equation specifies how h_t evolves; it does not specify which observable the observer chooses to measure, or what vocabulary the Born layer projects onto. The observer and the observable must be specified *in addition to* the dynamics, not derived from them. This is the same logical structure as in a transformer: the residual stream dynamics do not determine the unembedding matrix.

The many-worlds interpretation, the pilot-wave interpretation, and other alternatives all attempt to derive the Born layer from the Schrödinger layer—to eliminate the observer as a primitive. The Copenhagen axiom system says: do not attempt this. The observer is primitive. The interface is primitive. They are defined by their relationship to the engine, not by reduction to it.

3.6 A concrete example: the qubit

The simplest quantum system is a qubit: $h_t \in \mathbb{C}^2$. It makes both layers fully concrete.

The Fock layer

A qubit state is any unit vector in $\mathbb{C}^2$:

$$h = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad |\alpha|^2 + |\beta|^2 = 1. \quad (3.21)$$

Two distinguished basis states: $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. A general state $h = \alpha|0\rangle + \beta|1\rangle$ is a superposition—not “both at once” but a single vector with complex amplitudes assigned to both basis directions.

A Hamiltonian. Let $H = \frac{\omega}{2}\sigma_z$ with $\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. The eigenvalues are $E_{\pm} = \pm\omega/2$; the eigenstates are $|0\rangle$ and $|1\rangle$. The evolution:

$$h_t = e^{-i\omega t \sigma_z / 2} h_0 = \begin{pmatrix} e^{-i\omega t / 2} \alpha_0 \\ e^{i\omega t / 2} \beta_0 \end{pmatrix}. \quad (3.22)$$

Each component rotates at frequency $\omega/2$ in the complex plane, in opposite directions. The hidden state precesses.

The Bloch sphere. Up to a global phase, every qubit state can be written as

$$|h\rangle = \cos \frac{\theta}{2} |0\rangle + e^{i\phi} \sin \frac{\theta}{2} |1\rangle, \quad (3.23)$$

which defines a point on the unit sphere in $\mathbb{R}^3$ via the Bloch vector $\vec{r} = (\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta)$. The Schrödinger evolution with $H = \frac{\omega}{2} \sigma_z$ rotates the Bloch vector around the z -axis at angular speed ω . The RNN hidden state literally precesses like a compass needle.

The Born layer: multi-lingual readout

The same rotating qubit, read in different languages:

Language 1: measure σ_z . Eigenvectors $|0\rangle, |1\rangle$ (eigenvalues $+1, -1$). Born probabilities:

$$p(+1) = |\alpha_0|^2, \quad p(-1) = |\beta_0|^2. \quad (3.24)$$

These are constant in time—the σ_z eigenstates are the energy eigenstates, so the Born distribution does not change. If the qubit starts in $|0\rangle$, the Born distribution is a delta at $+1$: sampling is deterministic and trimming does nothing.

Language 2: measure σ_x . Eigenvectors $|\pm\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \pm 1 \end{pmatrix}$ (eigenvalues ± 1). Start from $h_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$:

$$p_t(+1) = \cos^2 \frac{\omega t}{2}, \quad p_t(-1) = \sin^2 \frac{\omega t}{2}. \quad (3.25)$$

The Born probabilities oscillate sinusoidally. This is *quantum interference*: it arises because the hidden state is complex-valued and rotating, and the probing vectors of σ_x are not aligned with the rotation axis.

Same vector, different languages, different distributions. The Fock layer is identical in both cases—the same h_t , the same rotation. The Born layer produces different distributions because the observer chose different vocabularies. This is basis agnosticism made concrete: the hidden state is not “in” the σ_z basis or the σ_x basis; it is a vector that can be read in either language.

Schrödinger versus Heisenberg picture

There are two equivalent ways to run the engine:

	Schrödinger picture	Heisenberg picture
Hidden state	Rotates: $h_t = e^{-iHt} h_0$	Fixed at h_0
Observable	Fixed	Rotates: $\hat{\mathcal{O}}_t = e^{iHt} \hat{\mathcal{O}} e^{-iHt}$

Both give identical Born probabilities. The Schrödinger picture says: the vector rotates, the probing vectors stand still. The Heisenberg picture says: the vector stands still, the probing vectors rotate (in the opposite direction). The relative angle is the same, so the inner products—and therefore the Born probabilities—are identical.

ML–QM correspondence table (updated)

ML / CS concept	QM concept (Fock-Born)
Hidden state $h_t \in \mathbb{C}^n$	Fock vector $ \psi(t)\rangle \in \mathcal{H}$
Linear RNN update $(I + W dt)h_t$	Schrödinger step $(I - iH dt) \psi\rangle$
Recurrent weight W	$-iH$ (Hamiltonian is the motor)
Rotation (unitary step)	e^{-iHdt} : norm-preserving evolution
Fixed output layer	Interactive Born layer: observer chooses observable
Single vocabulary	Multi-lingual: position, momentum, energy, spin, ...
Exp-softmax	Aggregated squared softmax (Born rule)
Partition function Z	1 (automatic from unitarity + completeness)
Output token	Eigenvalue λ (classical outcome)
Token embedding / input layer	Bohr trimming: $ \psi\rangle \leftarrow P_\lambda \psi\rangle / \ P_\lambda \psi\rangle\ $
Hard embedding lookup	Trimming in non-degenerate case: $ \psi\rangle \leftarrow \lambda\rangle$
Tokeniser (fixed vocabulary)	Choice of observable (eigenvalue vocabulary)
No output at layer 40	No classical variable between measurements
Screen dark between frames	Classical world absent between Born renderings
User outside the model	Observer outside $ \psi(t)\rangle$ (primitive)
Game engine	Fock layer: $h_t = e^{-iHt}h_0$
Rendering interface	Born layer: probability, sampling, trimming

4 Hamiltonians, Quantization, and Observables

The previous section introduced the architecture of quantum mechanics as an RNN. This section makes the architecture concrete. We specify what the Hamiltonian matrix H actually is — first classically, then quantum mechanically — and we give a precise account of observables: the observer’s choice of what to measure, which simultaneously determines the vocabulary of possible classical outputs (the tokenizer) and the unembedding matrix (the Born readout basis).

4.1 The classical Hamiltonian: total energy

Classically, the Hamiltonian H is simply the total energy of the system: kinetic plus potential. Before we can quantize, we need to know what H looks like as a function of position x and momentum p .

Free particle. A particle of mass $m = 1$ moving in one dimension with momentum p has only kinetic energy:

$$H = \frac{p^2}{2}. \quad (4.1)$$

Harmonic oscillator. Adding a restoring spring with unit spring constant gives potential energy $V(x) = \frac{1}{2}x^2$:

$$H = \frac{1}{2}(p^2 + x^2). \quad (4.2)$$

The harmonic oscillator is the most important system in physics. Any smooth potential near a local minimum looks quadratic by Taylor expansion, so the harmonic oscillator is the universal local approximation. It models molecular vibrations, photons in a cavity, phonons in a crystal, and is the building block of quantum field theory. Its role in physics is the same as the linear layer in neural networks: the elementary unit from which everything else is built.

General structure. In classical mechanics, x and p are real numbers, and $H(x, p)$ is a real-valued function. Hamilton’s equations:

$$\dot{x} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial x}, \quad (4.3)$$

govern the time evolution of the classical state $(x(t), p(t))$. Quantization replaces this classical evolution with the Schrödinger RNN.

4.2 Quantization: promoting scalars to operators

The quantization step replaces the classical observables x and p — real numbers — with operators (matrices) $\hat{X}$ and $\hat{P}$ acting on the Hilbert space $\mathcal{H}$. Every classical observable $f(x, p)$ becomes the corresponding operator $f(\hat{X}, \hat{P})$. The replacement is governed by one constraint: the *canonical commutation relation*.

The position operator $\hat{X}$. In the position basis $\{|x\rangle\}$, the position operator acts by multiplication:

$$\hat{X} |x\rangle = x |x\rangle, \quad (\hat{X}\psi)(x) = x \cdot \psi(x). \quad (4.4)$$

The position eigenstates $|x\rangle$ are the basis vectors; the eigenvalue is x . In this basis, $\hat{X}$ is diagonal: it is the “identity” of the position basis, acting on each basis vector by returning its label.

The momentum operator $\hat{P}$. In the position basis, the momentum operator acts as a derivative:

$$\hat{P} = -i \frac{d}{dx}, \quad (\hat{P}\psi)(x) = -i\psi'(x). \quad (4.5)$$

In Dirac notation:

$$\langle x | \hat{P} | \psi \rangle = -i \frac{d}{dx} \langle x | \psi \rangle = -i \frac{d}{dx} \psi(x). \quad (4.6)$$

The momentum operator is *not* diagonal in the position basis: it connects each position to its neighbours (it is a differential operator, or in the discrete approximation, a finite-difference matrix).

Discrete approximation. Discretise position: $x \in \{x_0, x_1, \dots, x_{n-1}\}$ with spacing s . The Hilbert space becomes $\mathbb{C}^n$ with basis vectors $|x_k\rangle$. The wavefunction $\psi(x_k) = \langle x_k | \psi \rangle$ is a column vector in $\mathbb{C}^n$. In this basis:

$$\hat{X} = \text{diag}(x_0, x_1, \dots, x_{n-1}), \quad \hat{P} \approx \frac{-i}{s} \begin{pmatrix} 1 & 0 & 0 & \dots \\ -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \dots \\ \vdots & & \ddots & \ddots \end{pmatrix}. \quad (4.7)$$

The position operator is diagonal (each basis state has a definite position); the momentum operator is a banded difference matrix (it mixes neighbouring position states). As $s \rightarrow 0$ and $n \rightarrow \infty$, these converge to the continuous operators (4.4) and (4.5).

Eigenvectors of $\hat{P}$: plane waves. The momentum eigenstates are plane waves. In the position basis:

$$\langle x | p \rangle = \frac{e^{ipx}}{\sqrt{2\pi}}, \quad \hat{P} | p \rangle = p | p \rangle. \quad (4.8)$$

A particle with definite momentum p is in state $|p\rangle$, whose position-basis wavefunction $\langle x | p \rangle = e^{ipx}/\sqrt{2\pi}$ is spread uniformly over all positions — the hallmark of momentum certainty and position uncertainty. In Dirac notation, the momentum basis is related to the position basis by the Fourier transform:

$$|p\rangle = \int dx \frac{e^{ipx}}{\sqrt{2\pi}} |x\rangle. \quad (4.9)$$

Remark. [The Fourier transform is a basis change] The Fourier transform is exactly the change-of-basis matrix from the position eigenbasis $\{|x\rangle\}$ to the momentum eigenbasis $\{|p\rangle\}$. The wavefunction in position space $\psi(x) = \langle x | \psi \rangle$ and the wavefunction in momentum space $\tilde{\psi}(p) = \langle p | \psi \rangle$ are the same state $|\psi\rangle$ expressed in two different bases. The Fourier transform converts between them. If you have ever used an FFT, you have been computing a quantum mechanical basis change.

4.3 The canonical commutation relation

Classically, x and p are numbers and $xp = px$. Quantum mechanically, $\hat{X}$ and $\hat{P}$ are matrices and they do not commute. The single equation:

$$[\hat{X}, \hat{P}] \equiv \hat{X}\hat{P} - \hat{P}\hat{X} = i\mathbf{1} \quad (4.10)$$

is the entire content of canonical quantization. Every consequence of quantum mechanics — the uncertainty principle, the energy levels of the harmonic oscillator, the hydrogen atom spectrum — follows from (4.10) together with the three-layer architecture.

Verification in the position basis. Apply $[\hat{X}, \hat{P}]$ to any wavefunction $\psi(x) = \langle x | \psi \rangle$:

$$(\hat{X}\hat{P}\psi)(x) = x \cdot (-i\psi'(x)) = -ix\psi'(x), \quad (4.11)$$

$$(\hat{P}\hat{X}\psi)(x) = -i \frac{d}{dx} (x\psi(x)) = -i\psi(x) - ix\psi'(x). \quad (4.12)$$

Subtracting:

$$([\hat{X}, \hat{P}]\psi)(x) = -ix\psi'(x) - (-i\psi(x) - ix\psi'(x)) = i\psi(x). \quad \checkmark \quad (4.13)$$

Non-commutativity in 2×2 : a toy model. In $\mathbb{C}^2$, take $\hat{X} = \sigma_x$ and $\hat{P} = \sigma_y$:

$$\sigma_x \sigma_y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \sigma_y \sigma_x = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}, \quad (4.14)$$

$$[\sigma_x, \sigma_y] = \begin{pmatrix} 2i & 0 \\ 0 & -2i \end{pmatrix}. \quad (4.15)$$

This does not equal $i\mathbf{1}$ exactly (the exact relation requires infinite dimensions), but it shows concretely that position and momentum matrices genuinely fail to commute.

The quantum Hamiltonian. Replacing $x \rightarrow \hat{X}$ and $p \rightarrow \hat{P}$ in the classical Hamiltonian:

$$H = \frac{1}{2}(\hat{P}^2 + \hat{X}^2) = \frac{1}{2}\left(-\frac{d^2}{dx^2} + x^2\right) \quad (\text{harmonic oscillator}). \quad (4.16)$$

In the discrete position basis (4.7), H is a concrete Hermitian matrix: the sum of the square of the difference matrix ($\hat{P}^2$, a tridiagonal matrix) and the square of the diagonal position matrix ($\hat{X}^2$). The Schrödinger RNN (5.1) with this H rotates the wavefunction vector $|\psi(t)\rangle \in \mathbb{C}^n$ forward in time.

4.4 Observables as the observer's choice: vocabulary, probing vectors, and degeneracy

An *observable* is a Hermitian operator $\hat{O}$ that the observer chooses to measure. This choice is not determined by the dynamics; it is not derivable from the Schrödinger equation. It is an independent decision made by the observer—the quantum analogue of choosing a tokeniser and an unembedding matrix simultaneously.

What the observable defines. The spectral decomposition of $\hat{O}$:

$$\hat{O} = \sum_{\lambda} \lambda P_{\lambda}, \quad P_{\lambda} = \sum_{k=1}^{d_{\lambda}} |\lambda, k\rangle \langle \lambda, k|, \quad (4.17)$$

simultaneously provides three things:

1. **The vocabulary** $\{\lambda\}$: the eigenvalues of $\hat{O}$, which are the only values that can appear as measurement outcomes. These are the classical words the Born layer can output—the possible readings on the observer's dial.
2. **The probing vectors** $\{|\lambda, k\rangle\}$: the eigenvectors of $\hat{O}$. Each probing vector is a direction in Hilbert space against which the hidden state is compared. The inner product $\langle \lambda, k | h_t \rangle$ is the “attention score” between the hidden state and the k -th probing vector for eigenvalue λ .
3. **The degeneracy** d_{λ} : the number of probing vectors that share the same eigenvalue λ . When $d_{\lambda} = 1$, the eigenvalue names a unique direction. When $d_{\lambda} > 1$, multiple orthogonal directions in Hilbert space produce the same classical output—the observable cannot distinguish them.

The Born probability for eigenvalue λ sums over all its probing vectors:

$$p(\lambda) = \sum_{k=1}^{d_{\lambda}} |\langle \lambda, k | h_t \rangle|^2 = \|P_{\lambda} h_t\|^2. \quad (4.18)$$

When $d_\lambda = 1$: $p(\lambda) = |\langle \lambda | h_t \rangle|^2$ —a single squared inner product, the textbook Born rule. When $d_\lambda > 1$: the Born probability is an OR sum over all probing vectors sharing the same eigenvalue. Cross terms between different k 's vanish by orthogonality—no interference within an eigensubspace.

The Bohr trimming keeps the projected superposition:

$$h_{t+} = \frac{P_\lambda h_t}{\|P_\lambda h_t\|} = \frac{\sum_{k=1}^{d_\lambda} |\lambda, k\rangle \langle \lambda, k | h_t \rangle}{\sqrt{p(\lambda)}}. \quad (4.19)$$

If $d_\lambda = 1$: the hidden state collapses to a single eigenvector $h_{t+} = |\lambda\rangle$ —complete trimming, complete information. If $d_\lambda > 1$: the hidden state is projected into the d_λ -dimensional eigensubspace, retaining its internal structure—partial trimming, partial information. The unresolved degrees of freedom (the choice of k within the eigensubspace) survive the measurement.

Non-degenerate case: the spectral decomposition simplifies. When every eigenvalue is non-degenerate ($d_\lambda = 1$ for all λ), the projector is rank-1: $P_\lambda = |\lambda\rangle \langle \lambda|$, and the spectral decomposition becomes:

$$\hat{\mathcal{O}} = \sum_{\lambda} \lambda |\lambda\rangle \langle \lambda|. \quad (4.20)$$

Each eigenvalue names one probing vector; there is no degeneracy. This is the textbook case, and it describes observables like position $\hat{x}$, momentum $\hat{p}$, and spin- z $\hat{S}_z$.

Most of this section uses the non-degenerate case for clarity. The degenerate case—where the observable's vocabulary is coarser than the Hilbert space dimension—is the subject of Section ??, and it is where temperature, decoherence, and classical physics emerge.

Two different tokenisers: position versus momentum. The choice between measuring position and measuring momentum is the choice between two different observables, two different vocabularies, two different sets of probing vectors, and two different Born distributions. The hidden state $|\psi(t)\rangle$ is the same in both cases; what changes is the read head.

Position measurement. Observable: $\hat{\mathcal{O}} = \hat{X}$, with spectral decomposition $\hat{X} = \int dx \, x |x\rangle \langle x|$.

Vocabulary:	$\{x : x \in \mathbb{R}\}$ (all possible positions).
Probing vectors:	$\{ x\rangle\}$ (position eigenstates = delta functions; one per eigenvalue, $d_x = 1$).
Born probability:	$p(x) = \langle x \psi \rangle ^2 = \psi(x) ^2$.
Bohr collapse:	$ \psi\rangle \leftarrow x_o\rangle$ (reset to the observed position).

Momentum measurement. Observable: $\hat{\mathcal{O}} = \hat{P}$, with spectral decomposition $\hat{P} = \int dp \, p |p\rangle \langle p|$.

Vocabulary:	$\{p : p \in \mathbb{R}\}$ (all possible momenta).
Probing vectors:	$\{ p\rangle\}$ (momentum eigenstates = plane waves; one per eigenvalue, $d_p = 1$).
Born probability:	$p(p) = \langle p \psi \rangle ^2 = \tilde{\psi}(p) ^2$.
Bohr collapse:	$ \psi\rangle \leftarrow p_o\rangle$ (reset to the observed momentum).

The same hidden state $|\psi(t)\rangle$ produces a different probability distribution and a different classical output depending on which observable is chosen. The position tokeniser and the momentum tokeniser are incompatible: choosing one makes the other maximally uncertain (this is the uncertainty principle, derived below).

	Position measurement	Momentum measurement
Observable $\hat{\mathcal{O}}$	$\hat{X} = \int x x\rangle \langle x dx$	$\hat{P} = \int p p\rangle \langle p dp$
Vocabulary (tokens)	$\{x \in \mathbb{R}\}$	$\{p \in \mathbb{R}\}$
Probing vectors	$\{ x\rangle\}$ (delta functions)	$\{ p\rangle\}$ (plane waves)
Born probability	$ \langle x \psi\rangle ^2 = \psi(x) ^2$	$ \langle p \psi\rangle ^2 = \tilde{\psi}(p) ^2$
Bohr collapse	$ \psi\rangle \leftarrow x_o\rangle$	$ \psi\rangle \leftarrow p_o\rangle$
Basis in position space	$\delta(x - x_o)$ (spike)	$e^{ip_o x} / \sqrt{2\pi}$ (sinusoid)

4.5 Energy as observable: the Hamiltonian’s dual role

The Hamiltonian H plays a unique dual role in the Fock-Born architecture. It is simultaneously:

1. **The motor of the Fock layer:** the operator that drives time evolution, $\dot{h}_t = -iHh_t$. This is its role inside the engine. The motor is always running; it does not require an observer.
2. **An observable of the Born layer:** a Hermitian operator the observer can choose to measure. Its eigenvalues $\{E_n\}$ are the energy vocabulary; its eigenstates $\{|E_n\rangle\}$ are the energy probing vectors. This is its role at the interface. It requires an observer who submits H to the Born protocol.

No other observable has this dual role. Position $\hat{x}$ is a Born-layer observable but not a motor. Momentum $\hat{p}$ is a Born-layer observable but not a motor. Spin $\hat{S}_z$ is a Born-layer observable but not a motor. The Hamiltonian is the only operator that appears on both sides of the architecture.

What “energy” means in the quantum framework. Classically, energy is a number: the total kinetic plus potential energy $E = p^2/2m + V(x)$ of a particle at definite position x with definite momentum p . In the quantum framework, energy undergoes the same transformation as every other classical quantity: it becomes an observable with a vocabulary and probing vectors.

The spectral decomposition of H :

$$H = \sum_n E_n |E_n\rangle \langle E_n| \quad (4.21)$$

(in the non-degenerate case) defines:

Vocabulary:	$\{E_0, E_1, E_2, \dots\}$ —the energy levels. These are the only values an energy measurement can return.
Probing vectors:	$\{ E_0\rangle, E_1\rangle, E_2\rangle, \dots\}$ —the energy eigenstates. Each is a direction in Hilbert space corresponding to a definite energy level.
Born probability:	$p(E_n) = \langle E_n h_t\rangle ^2$ —the probability of finding the system at energy level E_n .
Bohr collapse:	If energy E_n is observed, the hidden state is trimmed to $ E_n\rangle$.

Between measurements, the hidden state h_t is typically a superposition of many energy eigenstates: $h_t = \sum_n c_n e^{-iE_n t} |E_n\rangle$. The system does not “have” a definite energy—it has a Born distribution over energies, just as it has a Born distribution over positions and momenta. The energy is a Born-layer rendering, not a Fock-layer property. (This point is worth emphasising: the temptation to say “the system has energy E ” is a conflation. The system has a hidden state h_t ; the energy is what the Born layer returns when the observer submits H .)

The dual role: motor and observable. In the energy eigenbasis, the dual role becomes transparent. As the motor:

$$h_t = e^{-iHt}h_0 = \sum_n c_n e^{-iE_n t} |E_n\rangle. \quad (4.22)$$

Each energy eigenstate $|E_n\rangle$ rotates at angular frequency E_n (in units where $\hbar = 1$). The energy levels are the *rotation rates* of the engine: E_n is how fast the component along $|E_n\rangle$ phase-rotates in the complex plane. The motor is a collection of independent clocks, one per energy eigenstate, each ticking at its own frequency.

As the observable: a Born measurement of H asks, “at which frequency is the engine ticking?” The Born probability $p(E_n) = |c_n|^2$ is the fraction of the engine’s amplitude that is rotating at frequency E_n . The outcome E_n is not a property of the engine’s state in the Fock layer—it is a rendering through the Born layer, produced by projecting h_t onto $|E_n\rangle$ and squaring.

What makes the energy basis special. In the energy basis, the amplitudes $c_n(t)$ have a particularly simple time dependence: each acquires a phase $e^{-iE_n t}$ and nothing else. The magnitudes $|c_n|^2$ are constant in time. This means:

The Born probability for energy is *time-independent*: $p(E_n) = |c_n|^2$ does not change as the engine runs. If the system starts with 30% of its amplitude at E_1 and 70% at E_2 , these fractions persist forever (under the free evolution; interactions change them).

This is *not* true for other observables. The Born probability for position $p(x, t) = |\psi(x, t)|^2$ typically oscillates in time (the wavefunction spreads, sloshes, interferes). The Born probability for momentum similarly oscillates. Only the energy probabilities are stationary under the free evolution.

In ML terms: the energy basis is the basis in which the RNN weight matrix $e^{-iH\epsilon}$ is diagonal. The energy eigenstates are the *principal components* of the dynamics—the directions that do not mix under the engine’s evolution. Every other basis (position, momentum) mixes under time evolution, producing oscillating Born distributions. The energy basis is the one where the dynamics is trivial: each component just phase-rotates.

Discrete energy levels and the word “quantum.” For a free particle, the energy eigenvalues form a continuous band: $E = |\mathbf{k}|^2/2m$ for any momentum $\mathbf{k} \in \mathbb{R}^3$. The vocabulary is continuous, and the energy Born distribution can be spread over an entire interval.

For a bound system—a particle confined by a potential well, an electron in a hydrogen atom, a mass on a spring—the boundary conditions force the eigenvalues to be discrete:

$$E_0 < E_1 < E_2 < \dots \quad (4.23)$$

The vocabulary is a countable set of isolated points. Between consecutive levels, no energy value is allowed—the Born layer cannot render it, because there is no probing vector there. The energy is *quantised*: it comes in discrete amounts.

This is the *quantum* in quantum mechanics: the energy vocabulary is discrete for bound systems. The discreteness is not put in by hand; it is forced by the boundary conditions on the eigenstates of H . Atoms emit light only at specific frequencies because the photon frequency $\nu = (E_m - E_n)/h$ is the difference of two vocabulary entries, and the vocabulary is discrete.

Degenerate energy levels. An energy eigenvalue E_n may have degeneracy $g_n > 1$: there are g_n linearly independent eigenstates $|E_n, k\rangle$ ($k = 1, \dots, g_n$) sharing the same energy. This happens when the Hamiltonian has a symmetry: a unitary operator U that commutes with H

($[U, H] = 0$) can transform one eigenstate into another without changing the energy. The g_n degenerate states are related by the symmetry and form a representation of the symmetry group.

When energy is measured with degeneracy $g_n > 1$:

$$p(E_n) = \sum_{k=1}^{g_n} |\langle E_n, k | h_t \rangle|^2. \quad (4.24)$$

The Born probability sums over all g_n probing vectors—the observer learns the energy but not which degenerate state the system is in. The post-measurement state is a superposition within the g_n -dimensional eigensubspace, with the original amplitudes preserved.

The classic example: the hydrogen atom at energy level n has degeneracy $g_n = n^2$ (before fine structure), corresponding to the n^2 states with different angular momentum quantum numbers (ℓ, m) but the same energy. An energy measurement returns $E_n = -13.6 \text{ eV}/n^2$ but does not determine ℓ or m . To learn those, the observer must submit additional observables ($\hat{L}^2$ and $\hat{L}_3$) to the Born layer.

The Hamiltonian is not drama. The dual role of H creates a subtle temptation to conflate the two layers. Because the energy eigenstates are the motor’s principal directions, and the energy eigenvalues are the motor’s rotation rates, it is tempting to say “the system is at energy E_n ” as if this were a Fock-layer fact. It is not. In the Fock layer, the system is in a definite state h_t —a single vector in Hilbert space, typically a superposition of many energy eigenstates, rotating under e^{-iHt} . The energy E_n is what appears when the Born layer is invoked with observable H . Even the energy—the most “intrinsic” seeming of all physical quantities—is a Born-layer rendering, not a property of the engine’s hidden state.

4.6 Observables are Hermitian: eigenvalues are real

The requirement that observables be Hermitian ($\hat{O} = \hat{O}^\dagger$) is not an arbitrary convention. It has two consequences that are physically necessary, and both can be proved in a few lines.

Real eigenvalues. *Claim:* if $\hat{O} = \hat{O}^\dagger$ and $\hat{O}|\lambda\rangle = \lambda|\lambda\rangle$, then $\lambda \in \mathbb{R}$.

Proof: Take the inner product of the eigenvalue equation with $\langle\lambda|$:

$$\langle\lambda| \hat{O} |\lambda\rangle = \lambda \langle\lambda|\lambda\rangle = \lambda. \quad (4.25)$$

Now take the adjoint of the eigenvalue equation: $\langle\lambda| \hat{O}^\dagger = \lambda^* \langle\lambda|$. Using $\hat{O}^\dagger = \hat{O}$:

$$\langle\lambda| \hat{O} |\lambda\rangle = \lambda^* \langle\lambda|\lambda\rangle = \lambda^*. \quad (4.26)$$

Comparing (4.25) and (4.26): $\lambda = \lambda^*$, so $\lambda \in \mathbb{R}$. □

This is required because the vocabulary $\{\lambda\}$ consists of classical measurement outcomes—real numbers on a dial, a detector count, a pointer position. A complex eigenvalue would correspond to a complex-valued classical outcome, which has no physical meaning.

Orthogonal eigenstates. *Claim:* if $\hat{O}|\lambda\rangle = \lambda|\lambda\rangle$ and $\hat{O}|\mu\rangle = \mu|\mu\rangle$ with $\lambda \neq \mu$, then $\langle\lambda|\mu\rangle = 0$.

Proof: From the eigenvalue equations: $\langle\lambda| \hat{O} |\mu\rangle = \mu \langle\lambda|\mu\rangle$ and $\langle\lambda| \hat{O} |\mu\rangle = (\hat{O}^\dagger |\lambda\rangle)^\dagger |\mu\rangle = \lambda^* \langle\lambda|\mu\rangle = \lambda \langle\lambda|\mu\rangle$ (using $\lambda^* = \lambda$). Subtracting: $(\lambda - \mu) \langle\lambda|\mu\rangle = 0$. Since $\lambda \neq \mu$: $\langle\lambda|\mu\rangle = 0$. □

Orthogonality is required for the Born probabilities to sum to 1: if the probing vectors $\{|\lambda\rangle\}$ were not orthonormal, the completeness relation $\sum_\lambda |\lambda\rangle \langle\lambda| = \mathbf{1}$ would fail, and the automatic normalisation of the Born distribution would break down.

Concrete example: the qubit. In $\mathbb{C}^2$, the three Pauli matrices:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (4.27)$$

are all Hermitian with eigenvalues ± 1 (real, as proved above). Each defines a different observable—a different tokeniser with the same two-token vocabulary $\{+1, -1\}$ but different probing vectors:

$$\sigma_z: \quad | +1 \rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad | -1 \rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (\text{computational basis}), \quad (4.28)$$

$$\sigma_x: \quad | +1 \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad | -1 \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad (\text{Hadamard basis}), \quad (4.29)$$

$$\sigma_y: \quad | +1 \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}, \quad | -1 \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} \quad (\text{circular basis}). \quad (4.30)$$

The same hidden state $|\psi\rangle \in \mathbb{C}^2$ produces a different Born distribution depending on which Pauli is chosen as the observable. The vocabulary is always $\{+1, -1\}$, but the probing vectors change: different observables are different read heads. One can verify orthogonality: $\langle +1 | -1 \rangle = 0$ in each basis, as the theorem guarantees.

4.7 The uncertainty principle as incompatible tokenisers

Two observables $\hat{A}$ and $\hat{B}$ are *compatible* if they commute: $[\hat{A}, \hat{B}] = 0$. Compatible observables share a common eigenbasis—the same set of probing vectors serves both tokenisers. They can be measured simultaneously, and measuring one does not disturb the other.

Two observables are *incompatible* if $[\hat{A}, \hat{B}] \neq 0$. Their probing vectors point in different directions—the two tokenisers are different—and measuring one forces the other to be uncertain. The Heisenberg uncertainty principle quantifies this:

$$\boxed{\Delta A \cdot \Delta B \geq \frac{1}{2} |\langle [\hat{A}, \hat{B}] \rangle|}, \quad (4.31)$$

where $\Delta A = \sqrt{\langle \hat{A}^2 \rangle - \langle \hat{A} \rangle^2}$ is the Born standard deviation of $\hat{A}$ in the state h_t .

Derivation from the Cauchy–Schwarz inequality. Define the shifted operators $\hat{A}' = \hat{A} - \langle \hat{A} \rangle$, $\hat{B}' = \hat{B} - \langle \hat{B} \rangle$ (subtracting the means does not change the commutator: $[\hat{A}', \hat{B}'] = [\hat{A}, \hat{B}]$). Define two vectors in Hilbert space:

$$|f\rangle = \hat{A}' |h_t\rangle, \quad |g\rangle = \hat{B}' |h_t\rangle. \quad (4.32)$$

Their norms are the variances: $\langle f|f \rangle = \langle h_t | \hat{A}'^2 |h_t \rangle = (\Delta A)^2$ and $\langle g|g \rangle = (\Delta B)^2$. The Cauchy–Schwarz inequality says:

$$|\langle f|g \rangle|^2 \leq \langle f|f \rangle \langle g|g \rangle = (\Delta A)^2 (\Delta B)^2. \quad (4.33)$$

Now decompose $\langle f|g \rangle = \langle h_t | \hat{A}' \hat{B}' |h_t \rangle$ into its real and imaginary parts:

$$\hat{A}' \hat{B}' = \frac{1}{2} \{ \hat{A}', \hat{B}' \} + \frac{1}{2} [\hat{A}', \hat{B}'], \quad (4.34)$$

where $\{ \hat{A}', \hat{B}' \} = \hat{A}' \hat{B}' + \hat{B}' \hat{A}'$ is the anticommutator (Hermitian, so its expectation is real) and $[\hat{A}', \hat{B}']$ is the commutator (anti-Hermitian, so its expectation is purely imaginary). Therefore:

$$|\langle f|g \rangle|^2 = \frac{1}{4} |\langle \{ \hat{A}', \hat{B}' \} \rangle|^2 + \frac{1}{4} |\langle [\hat{A}', \hat{B}'] \rangle|^2 \geq \frac{1}{4} |\langle [\hat{A}, \hat{B}] \rangle|^2, \quad (4.35)$$

where we dropped the non-negative anticommutator term. Combining with (4.33):

$$(\Delta A)^2(\Delta B)^2 \geq \frac{1}{4} |\langle [\hat{A}, \hat{B}] \rangle|^2. \quad (4.36)$$

Taking the square root gives the uncertainty principle (4.31). $\square$

Application: position and momentum. For $\hat{A} = \hat{X}$ and $\hat{B} = \hat{P}$: $[\hat{X}, \hat{P}] = i\mathbf{1}$, so $|\langle [\hat{X}, \hat{P}] \rangle| = |\langle i\mathbf{1} \rangle| = 1$, giving:

$$\Delta x \cdot \Delta p \geq \frac{1}{2}. \quad (4.37)$$

No state in the Hilbert space can violate this bound. It is saturated by the coherent states of the harmonic oscillator (Section 6.8).

In the tokeniser language. The uncertainty principle is a change-of-basis effect: a hidden state that is one-hot in one tokeniser's probing vectors (certainty in $\hat{A}$) must be dense in an incompatible tokeniser's probing vectors (uncertainty in $\hat{B}$). If $|h_t\rangle$ is an eigenstate of $\hat{A}$ (perfectly aligned with one of $\hat{A}$'s probing vectors), then expanding $|h_t\rangle$ in $\hat{B}$'s probing vectors $\{|b\rangle\}$:

$$|h_t\rangle = \sum_b \langle b|h_t\rangle |b\rangle, \quad (4.38)$$

the coefficients $\langle b|h_t\rangle$ cannot all vanish except one (because $[\hat{A}, \hat{B}] \neq 0$ means the two sets of probing vectors are rotated with respect to each other). The Born distribution for $\hat{B}$ is forced to be spread across multiple eigenvalues. Certainty in $\hat{A}$'s vocabulary forces uncertainty in $\hat{B}$'s vocabulary.

Concrete example: σ_z and σ_x on a qubit. Prepare the hidden state $|h\rangle = |0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, the $+1$ eigenstate of σ_z . Measuring σ_z gives outcome $+1$ with certainty:

$$p_{\sigma_z}(+1) = |\langle +1_z|h\rangle|^2 = \left| \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|^2 = 1. \quad (4.39)$$

Now measure σ_x instead. Projecting onto σ_x 's probing vectors:

$$p_{\sigma_x}(+1) = |\langle +1_x|h\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|^2 = \frac{1}{2}, \quad (4.40)$$

$$p_{\sigma_x}(-1) = |\langle -1_x|h\rangle|^2 = \left| \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right|^2 = \frac{1}{2}. \quad (4.41)$$

Knowing $\sigma_z = +1$ with certainty forces σ_x to be maximally random: equal probability of $+1$ and -1 . The two tokenisers are maximally incompatible—their probing vectors are rotated by 45° in Hilbert space.

5 The Schrödinger equation: three formulations and one theory

The Schrödinger equation in Dirac notation:

$$i\hbar \frac{d}{dt} |h_t\rangle = H |h_t\rangle \quad (5.1)$$

is basis-agnostic. It says: the hidden state $|h_t\rangle$ rotates in Hilbert space, driven by the Hamiltonian H . No coordinates, no wave, no matrix, no wavefunction—just a vector and an operator. The vector $|h_t\rangle$ is not written in any particular basis; it is the abstract object itself, and it can be viewed from any basis the observer chooses. Projecting onto the position basis gives a PDE for $\psi(x, t)$. Projecting onto the energy basis gives a diagonal system of ODEs for $c_n(t)$. Projecting onto the momentum basis gives a PDE for $\tilde{\psi}(p, t)$. Projecting onto the Fock basis gives the QFT evolution of amplitude across particle configurations. All are the same equation (5.1), viewed through different lenses.

But this notation came last, not first. The equation was discovered three times in eighteen months, in three different languages, by three physicists who did not initially realise they were saying the same thing. Understanding all three formulations—and how Dirac unified them—is the best way to understand why the basis-agnostic notation is the right one, and why it is essential for quantum field theory.

5.1 Heisenberg’s matrix mechanics (1925): the commutator encodes the dynamics

In the summer of 1925, Werner Heisenberg, suffering from hay fever, retreated to the island of Helgoland in the North Sea. Working in isolation, he made a radical decision: to construct a mechanics that referred only to *observable* quantities—the frequencies and intensities of spectral lines—and not to unobservable electron orbits.

Heisenberg arranged the observable transition amplitudes into tables: one table for position, one for momentum. Each table had entries indexed by two quantum numbers (m, n) , representing the transition from state m to state n . He discovered a composition rule for these tables: to compute the table for XY (position times momentum), multiply the entries as:

$$(XY)_{mn} = \sum_k X_{mk} Y_{kn}. \quad (5.2)$$

This is matrix multiplication—but Heisenberg did not know it. What troubled him was that the composition was *not commutative*: $XY \neq YX$. He wrote in his paper that this feature was “very unpleasant” and that he had “not been able to circumvent” it.

It was Heisenberg’s mentor Max Born who, on reading the paper, recognized (5.2) as matrix multiplication. Born had studied matrix algebra as a student; Heisenberg had not. Together with Pascual Jordan, Born formalised the theory in the “Dreimännerarbeit” (three-man paper) of 1925. The key result: the commutation relation

$$XP - PX = i\hbar \mathbf{1} \quad (5.3)$$

is not an unpleasant defect—it is the *fundamental dynamical rule* of quantum mechanics. The commutator encodes the entire dynamics: Heisenberg’s equation of motion for any observable A is:

$$i\hbar \frac{dA}{dt} = [A, H] = AH - HA. \quad (5.4)$$

The observable A evolves in time; the state is fixed. This is the *Heisenberg picture*: the matrices move, the vector stays still.

What Heisenberg got right. Heisenberg’s formulation is closer to the final mathematical structure than Schrödinger’s, even though it is less familiar. The observables are operators (matrices); the physical content is in their commutation relations; the dynamics is $dA/dt = [A, H]/i\hbar$. The state vector $|h\rangle$ —which Heisenberg did not write down explicitly—is implicit: it is the vector on which the matrices act. The matrix elements $A_{mn} = \langle m|A|n\rangle$ are the transition amplitudes between energy eigenstates—the physically observable quantities that Heisenberg started from.

5.2 Schrödinger’s wave mechanics (1926): the wave equation

In January 1926, Erwin Schrödinger published the first of four papers that would transform physics. His approach was opposite to Heisenberg’s: instead of starting from observables and discovering matrices, Schrödinger started from a *wave equation* for a function $\psi(x, t)$:

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x)\psi. \quad (5.5)$$

This is the Schrödinger wave equation: a partial differential equation for the complex-valued function $\psi(x, t)$. Schrödinger derived it by analogy with classical wave optics and the de Broglie relation $p = \hbar k$.

In the Dirac notation developed later, equation (5.5) is the position-basis representation of $i\hbar \partial_t |h_t\rangle = H |h_t\rangle$. To see this, project both sides of (5.1) onto $\langle x|$:

$$i\hbar \frac{\partial}{\partial t} \langle x|h_t\rangle = \langle x|H|h_t\rangle. \quad (5.6)$$

The left side is $i\hbar \partial_t \psi(x, t)$, since $\psi(x, t) = \langle x|h_t\rangle$. The right side, for $H = \hat{p}^2/2m + V(\hat{x})$ with $\langle x|\hat{p} = -i\hbar \partial_x \langle x|$, gives $[-\hbar^2 \partial_x^2/2m + V(x)]\psi(x, t)$: Schrödinger’s wave equation. The wave equation is the Dirac equation projected onto the position basis.

Schrödinger’s initial interpretation—and its correction. Schrödinger initially interpreted $\psi(x, t)$ as something physically real: a charge density $e|\psi(x)|^2$ smeared through space, like a classical wave. This interpretation was quickly shown to be untenable: for two particles, $\psi(x_1, x_2)$ is a function on $\mathbb{R}^6$, not a field in 3D space; it cannot be a physical wave in ordinary space.

Born corrected the interpretation in 1926: $|\psi(x)|^2$ is the *probability density* for finding the particle at x , not a charge density. In the Dirac language: $\psi(x) = \langle x|h_t\rangle$ is the coefficient of the position basis vector $|x\rangle$ in the expansion of the hidden state $|h_t\rangle$. The wavefunction is not the physical reality; it is the representation of the hidden state in a particular basis. The physical reality is $|h_t\rangle$, which exists independently of any basis.

Schrödinger’s equivalence proof. In his fourth paper (1926), Schrödinger proved that his wave mechanics and Heisenberg’s matrix mechanics give identical predictions for all observable quantities. The proof amounts to showing that the matrix elements $A_{mn} = \int dx \psi_m^*(x) \hat{A} \psi_n(x)$ computed from the wavefunctions equal Heisenberg’s table entries. This is the statement that $\langle m|A|n\rangle$ can be computed in any basis: in the energy basis (Heisenberg’s matrices) or the position basis (Schrödinger’s wavefunctions). The two formulations are the same theory in different bases.

The narrow and broad meanings of “the Schrödinger equation.”

The term “Schrödinger equation” is used in two senses throughout physics, and confusing them causes unnecessary difficulty:

The narrow sense: the wave equation (5.5), a PDE for $\psi(x, t)$ in the position basis. This is what Schrödinger wrote down in 1926. It is the position-basis projection $\langle x | i\hbar \partial_t | h_t \rangle = \langle x | H | h_t \rangle$ of the Dirac equation (5.1).

The broad sense: the Dirac equation $i\hbar d|h_t\rangle/dt = H|h_t\rangle$, a first-order ODE for the abstract vector $|h_t\rangle$ in Hilbert space. This is basis-agnostic: it does not mention x , p , or any particular representation.

The narrow-sense equation is convenient for single-particle quantum mechanics, where the position basis is natural and the Hilbert space is $L^2(\mathbb{R}^3)$. But it becomes progressively more inconvenient as the system grows:

For n particles, the wavefunction $\psi(x_1, \dots, x_n, t)$ lives on $\mathbb{R}^{3n}$ —a $3n$ -dimensional space. The PDE is a partial differential equation in $3n + 1$ variables. For $n \sim 10^{23}$ (a macroscopic system), this is a function of $\sim 10^{23}$ variables—utterly intractable.

For quantum field theory, the situation is worse: the number of particles is not fixed, so the “wavefunction” is an infinite tower of functions $\psi_0, \psi_1(x_1), \psi_2(x_1, x_2), \dots$, one for each particle-number sector. The position-basis PDE becomes an infinite coupled system of PDEs on spaces of increasing dimension. This is why the Schrödinger wave equation—in the narrow sense—almost never appears in quantum field theory textbooks.

The broad-sense equation $i\hbar d|h_t\rangle/dt = H|h_t\rangle$ suffers none of these problems. The Fock vector $|h_t\rangle$ is a single object in the Fock space $\mathcal{F} = \bigoplus_n \mathcal{H}^{(n)}$, encoding all particle numbers simultaneously as a superposition. The Hamiltonian H is a polynomial of creation and annihilation operators, acting on this single vector. The equation is a matrix-vector ODE—the same structure as a linear RNN—regardless of whether the system has 1 particle, 10^{23} particles, or a variable number. The Fock vector does not live in position space; it lives in the abstract Hilbert space, and it can be projected onto any basis as needed.

This is why Dirac notation is not merely convenient but *essential* for quantum field theory. The position-basis wavefunction $\psi(x, t)$ is a useful tool for single-particle problems. The Fock vector $|h_t\rangle$ is the right language for everything.

	Schrödinger (narrow)	Dirac (broad)
Object	$\psi(x_1, \dots, x_n, t) \in L^2(\mathbb{R}^{3n})$	$ h_t\rangle \in \mathcal{F}$
Equation	PDE in $3n + 1$ variables	ODE for a vector: $\dot{h}_t = -iHh_t$
Particle number	Fixed at n	Variable (superposition over n)
For $n = 1$	Natural	Also natural
For $n \sim 10^{23}$	Intractable	Same equation, larger Hilbert space
For variable n (QFT)	Infinite tower of PDEs	Same equation, Fock space
Basis	Position basis committed	Basis-agnostic

Throughout this book, “the Schrödinger equation” means the broad-sense Dirac equation (5.1): $\dot{h}_t = -iHh_t$, a first-order linear ODE for a vector in Hilbert space. The position-basis wave equation (5.5) is a projection of this equation onto a particular basis, useful for specific calculations but not the foundation.

5.3 Dirac’s q-numbers (1925–26): the logical leap

In the autumn of 1925, Heisenberg gave a talk at Cambridge. In the audience was Paul Dirac, a 23-year-old graduate student. Dirac did not follow the details of Heisenberg’s talk, but he

grasped the essential point: the dynamical variables of quantum mechanics do not commute, $AB \neq BA$.

What followed was one of the great logical leaps in physics. Dirac recognized that the commutator $[A, B] = AB - BA$ plays the same structural role in quantum mechanics that the Poisson bracket $\{A, B\}_{\text{PB}} = \sum_i (\partial A / \partial q_i \partial B / \partial p_i - \partial A / \partial p_i \partial B / \partial q_i)$ plays in classical mechanics. Classical mechanics says: $dA/dt = \{A, H\}_{\text{PB}}$. Quantum mechanics says: $dA/dt = [A, H]/i\hbar$. The correspondence is:

$$\{A, B\}_{\text{PB}} \longleftrightarrow \frac{1}{i\hbar}[A, B]. \quad (5.7)$$

The canonical Poisson bracket $\{q_i, p_j\}_{\text{PB}} = \delta_{ij}$ becomes $[q_i, p_j] = i\hbar\delta_{ij}$: the Born–Jordan commutation relation (5.3), derived by Dirac from a completely different starting point.

q-numbers. Dirac called the non-commuting dynamical variables *q-numbers* (quantum numbers), to distinguish them from ordinary *c-numbers* (classical numbers) that commute. He did not work with matrices or wavefunctions: he worked directly with the algebraic relations between q-numbers, deriving physical predictions from the commutation relations alone.

This was Dirac’s sixth sense at work. Where Heisenberg saw tables and Born saw matrices and Schrödinger saw wavefunctions, Dirac saw the *algebraic structure*: a set of objects (*q-numbers*) with a non-commutative product and a bracket related to the Poisson bracket. He did not need to specify what the q-numbers *were* (matrices? differential operators? something else?) to extract the physics. The physics was in the algebra, not in the representation.

This is the same insight that underlies basis agnosticism: the physical content is in the inner products and commutators, not in the specific components. Dirac arrived at this insight before inventing the notation to express it.

5.4 Dirac’s transformation theory (1927): the unification

Dirac’s discoveries often involve dazzling logical leaps—the Dirac equation, the prediction of the positron, the magnetic monopole argument. But his *transformation theory*, published in 1927, is different: it is absolutely logical, step by step, with no gap in the reasoning. Dirac called it his darling, and it is the work he was most proud of.

The problem transformation theory solves: Heisenberg’s matrix mechanics works in the energy basis $\{|E_n\rangle\}$; Schrödinger’s wave mechanics works in the position basis $\{|x\rangle\}$; and Dirac’s own q-number algebra works without any basis at all. How are these related? Is there a single mathematical framework that contains all three as special cases?

The framework. Dirac’s answer: quantum mechanics is a theory of a vector space $\mathcal{H}$ (the Hilbert space) equipped with an inner product $\langle\phi|\psi\rangle$. The physical content is entirely in the inner products. Every basis $\{|\alpha\rangle\}$ gives a *representation* of the theory: the state $|\psi\rangle$ is represented by its components $\langle\alpha|\psi\rangle$, and the operator A by its matrix elements $\langle\alpha|A|\beta\rangle$. Different bases give different representations of the same physics.

The transformation. A change of basis from $\{|i\rangle\}$ to $\{|\alpha\rangle\}$ is a unitary transformation U defined by:

$$U_{i\alpha} = \langle i|\alpha\rangle. \quad (5.8)$$

The components of a state transform as:

$$\psi_i = \langle i|\psi\rangle = \sum_{\alpha} \langle i|\alpha\rangle \langle\alpha|\psi\rangle = \sum_{\alpha} U_{i\alpha} \tilde{\psi}_{\alpha}. \quad (5.9)$$

The matrix elements of an operator transform as:

$$A_{ij} = \langle i | A | j \rangle = \sum_{\alpha, \beta} U_{i\alpha} \tilde{A}_{\alpha\beta} U_{j\beta}^*. \quad (5.10)$$

Both are obtained by inserting the resolution of the identity $\sum_{\alpha} |\alpha\rangle \langle \alpha| = \mathbf{1}$ —the same trick from the previous section. The transformation is unitary: $\sum_i U_{i\alpha} U_{i\beta}^* = \delta_{\alpha\beta}$ (from $\langle \alpha | \beta \rangle = \delta_{\alpha\beta}$).

The inner product is invariant. The inner product $\langle \phi | \psi \rangle$ does not change under the transformation:

$$\langle \phi | \psi \rangle = \sum_i \phi_i^* \psi_i = \sum_{\alpha} \tilde{\phi}_{\alpha}^* \tilde{\psi}_{\alpha}. \quad (5.11)$$

Both sums give the same number. The Born probability $P(\lambda) = |\langle \lambda | \psi \rangle|^2$, which is an inner product, is therefore the same in every representation. The expectation value $\langle A \rangle = \langle \psi | A | \psi \rangle$ is invariant. The eigenvalues of A are invariant (they are the roots of $\det(A - \lambda \mathbf{1}) = 0$, which is invariant under unitary transformation). Everything that can be measured is invariant.

Schrödinger and Heisenberg as two representations. The equivalence of wave mechanics and matrix mechanics is now immediate. Schrödinger's formulation uses the position basis $\{|x\rangle\}$: the state is $\psi(x) = \langle x | \psi \rangle$, and the Hamiltonian is the differential operator $H = -\hbar^2 \partial_x^2 / 2m + V(x)$. Heisenberg's formulation uses the energy basis $\{|E_n\rangle\}$: the state is $c_n = \langle E_n | \psi \rangle$, and the Hamiltonian is the diagonal matrix $H_{mn} = E_n \delta_{mn}$. The transformation between them is:

$$\psi(x) = \sum_n \langle x | E_n \rangle c_n = \sum_n \psi_n(x) c_n, \quad (5.12)$$

where $\psi_n(x) = \langle x | E_n \rangle$ is the n -th energy eigenfunction in the position basis. The transformation matrix $U_{xn} = \langle x | E_n \rangle = \psi_n(x)$ is unitary (the energy eigenfunctions form an orthonormal set in L^2). Schrödinger's equivalence proof is the statement that U is unitary: $\sum_n \psi_n^*(x) \psi_n(x') = \delta(x - x')$.

The momentum representation. The momentum basis $\{|p\rangle\}$ gives a third representation. The transformation from position to momentum is the Fourier transform:

$$\tilde{\psi}(p) = \langle p | \psi \rangle = \int dx \langle p | x \rangle \langle x | \psi \rangle = \int dx \frac{e^{-ipx/\hbar}}{\sqrt{2\pi\hbar}} \psi(x). \quad (5.13)$$

The transformation matrix $U_{px} = \langle p | x \rangle = e^{-ipx/\hbar} / \sqrt{2\pi\hbar}$ is the Fourier kernel. The position and momentum representations are related by a unitary transformation, and all physical predictions are identical in both.

The Schrödinger and Heisenberg pictures. Transformation theory also clarifies the two “pictures” of quantum dynamics.

In the *Schrödinger picture*, the state evolves and the operators are fixed:

$$|h_t\rangle = e^{-iHt/\hbar} |h_0\rangle, \quad A_S = A. \quad (5.14)$$

In the *Heisenberg picture*, the operators evolve and the state is fixed:

$$|h\rangle_H = |h_0\rangle, \quad A_H(t) = e^{iHt/\hbar} A e^{-iHt/\hbar}. \quad (5.15)$$

The two are related by the unitary transformation $U(t) = e^{-iHt/\hbar}$. All inner products $\langle h_t | A | h_t \rangle = \langle h_0 | A_H(t) | h_0 \rangle$ are identical. The Schrödinger picture rotates the vector; the Heisenberg picture

rotates the operators in the opposite direction. Both are representations of the same physics, related by a time-dependent change of basis.

In the RNN language: the Schrödinger picture evolves the hidden state h_t through the weight matrix $e^{-iH\Delta t}$ at each step. The Heisenberg picture keeps the hidden state fixed and instead transforms the unembedding head (the observable) at each step. Both compute the same Born probabilities.

5.5 Dirac’s Lagrangian paper (1933): the seed of the path integral

Transformation theory connects different bases of the same Hilbert space at a single time. But it also connects the *same* basis at *different* times: the propagator $\langle x_f, t_f | x_i, t_i \rangle$ is the matrix element of the time-evolution operator in the position basis, and it is a change-of-basis amplitude from the initial position to the final position.

In 1933, Dirac published a short paper—“The Lagrangian in Quantum Mechanics”—that asked a deceptively simple question: what is the form of this amplitude for an infinitesimal time step?

The question. The propagator for a finite time interval is:

$$\langle x_f, t_f | x_i, t_i \rangle = \langle x_f | e^{-iH(t_f - t_i)/\hbar} | x_i \rangle. \quad (5.16)$$

For an infinitesimal time step $\varepsilon = t_f - t_i$:

$$\langle x_f, t_i + \varepsilon | x_i, t_i \rangle = \langle x_f | e^{-iH\varepsilon/\hbar} | x_i \rangle. \quad (5.17)$$

In the matrix language of this book: this is the (x_f, x_i) matrix element of the one-step evolution matrix $U_\varepsilon = e^{-iH\varepsilon/\hbar}$.

Dirac’s observation. Dirac observed that this matrix element is “analogous to” $e^{iL\varepsilon/\hbar}$, where L is the classical Lagrangian evaluated along the classical path from x_i to x_f in time ε :

$$\langle x_f, t_i + \varepsilon | x_i, t_i \rangle \quad \text{“corresponds to”} \quad \exp\left(\frac{i}{\hbar} L\left(x_i, \frac{x_f - x_i}{\varepsilon}\right)\varepsilon\right). \quad (5.18)$$

The Lagrangian $L(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - V(x)$ is the classical kinetic minus potential energy, evaluated at position x_i with velocity $(x_f - x_i)/\varepsilon$.

Dirac was characteristically precise about the status of this relation. He did not write an equals sign; he wrote “analogous to.” He did not specify the proportionality constant. He did not take the next step of composing many infinitesimal steps into a finite-time propagator. He noted the correspondence, derived its consequences for the classical limit, and stopped.

The finite-time propagator. Had Dirac taken one more step, he would have invented the path integral. The finite-time propagator is obtained by inserting $N - 1$ resolutions of the identity (as in our Dirac notation section):

$$\begin{aligned} \langle x_f, t_f | x_i, t_i \rangle &= \int dx_1 \cdots dx_{N-1} \langle x_f | x_{N-1} \rangle_\varepsilon \cdots \langle x_1 | x_i \rangle_\varepsilon \\ &\quad \text{“corresponds to”} \quad \int dx_1 \cdots dx_{N-1} \exp\left(\frac{i}{\hbar} \sum_{k=0}^{N-1} L\left(x_k, \frac{x_{k+1} - x_k}{\varepsilon}\right)\varepsilon\right). \end{aligned} \quad (5.19)$$

The exponent is $iS[x(t)]/\hbar$, where $S = \int L dt$ is the classical action along the piecewise-linear path $(x_i, x_1, \dots, x_{N-1}, x_f)$. The integral over all intermediate positions is a sum over all paths.

This is the path integral:

$$\langle x_f, t_f | x_i, t_i \rangle = \int \mathcal{D}[x(t)] e^{iS[x(t)]/\hbar}. \quad (5.20)$$

The composition of infinitesimal propagators—each “analogous to” $e^{iL\varepsilon/\hbar}$ —is the matrix power formula (7.4) from the Dirac notation section, written in the position basis with continuous intermediate indices.

What Dirac did and did not do. Dirac’s 1933 paper contains every ingredient of the path integral: the infinitesimal propagator as $e^{iL\varepsilon/\hbar}$, the composition via intermediate-state insertions, the appearance of the classical action $S = \int L dt$ in the exponent. What it does not contain is the final synthesis: writing the finite-time propagator as a single functional integral over all paths, recognizing this as a new formulation of quantum mechanics, and developing the computational machinery (Feynman diagrams, the generating functional, perturbation theory in the path integral language).

Dirac saw the correspondence as a property of the existing formalism (transformation theory), not as a new formulation. He considered it an observation, not a programme.

Feynman’s completion (1942–48)

Richard Feynman, as a graduate student at Princeton, read Dirac’s 1933 paper (reportedly brought to his attention by Herbert Jehle at a beer party in 1941). Feynman asked whether Dirac’s “analogous to” could be made into an “equals.” He worked out the proportionality constant (the normalization $\sqrt{m/2\pi i\hbar\varepsilon}$ per intermediate integration) and showed that the composition (5.19), taken to the continuum limit $N \rightarrow \infty$, $\varepsilon \rightarrow 0$, gives the exact quantum propagator.

Feynman’s thesis (1942, supervised by John Wheeler) developed the path integral into a complete formulation of quantum mechanics. His subsequent papers (1948–49) extended it to QED, introduced Feynman diagrams as a computational tool for the perturbative expansion of the path integral, and showed that the results are equivalent to the operator formalism of Schwinger and Tomonaga.

What Feynman added to Dirac.

1. **The equals sign.** Dirac wrote “analogous to”; Feynman proved the equality, with the correct normalization.
2. **The functional integral.** Dirac composed finitely many propagators; Feynman took the continuum limit and defined the integral over all paths as a new mathematical object $\int \mathcal{D}[x(t)]$.
3. **A new formulation.** Dirac saw the Lagrangian correspondence as a property of transformation theory; Feynman made it an independent starting point for quantum mechanics, equivalent to but conceptually distinct from the Schrödinger and Heisenberg formulations.
4. **The classical limit.** In the limit $\hbar \rightarrow 0$, the phase $e^{iS/\hbar}$ oscillates wildly and the integral is dominated by the path of stationary phase: $\delta S = 0$, the classical path. The classical limit of the path integral is Hamilton’s principle of least action. Feynman made this connection precise; Dirac had noted it qualitatively.
5. **The extension to fields.** Feynman’s original path integral is over particle trajectories $x(t)$. The extension to the field path integral $\int \mathcal{D}[\phi] e^{iS[\phi]/\hbar}$ —integrating over all field configurations rather than all particle trajectories—came later and is the foundation of the log Z framework used throughout this book.

In the framework of this book. The path integral is not a separate formulation; it is the position-basis representation of the matrix power formula $\langle x_f | U^N | x_i \rangle = \int dx_1 \cdots dx_{N-1} \prod_k (U_\varepsilon)_{x_{k+1}, x_k}$, which we derived in the Dirac notation section from the resolution of the identity. Dirac’s 1933

observation identifies the matrix elements $(U_\varepsilon)_{x_{k+1}, x_k}$ with $e^{iL\varepsilon/\hbar}$. Feynman’s contribution was to recognize that this identification, composed over many time steps and taken to the continuum, is a complete and computationally powerful formulation of quantum mechanics.

The route from Dirac’s transformation theory (1927) through Dirac’s Lagrangian paper (1933) to Feynman’s path integral (1942–48) is a single logical arc: the matrix element of the time-evolution operator in the position basis, written as a product of infinitesimal matrix elements via the resolution of the identity, with each infinitesimal factor identified as $e^{iL\varepsilon/\hbar}$. The path integral is transformation theory applied to time evolution.

5.6 Three formulations, one vector, many lenses

We now have three ways to express the same physics:

Formulation	Basis	State is...	Dynamics is...	Best for...
Heisenberg	Energy $ E_n\rangle$	Fixed vector	Matrices evolve	Algebraic spectrum
Schrödinger	Position $ x\rangle$	PDE for $\psi(x, t)$	$-\nabla^2/2m + V(x)$	1-particle problems
Dirac	Any / none	Abstract $ h_t\rangle$	$\dot{h}_t = -iHh_t$	Everything
Path integral	Position $ x\rangle$, time-sliced	Sum over paths	$e^{iS[x]/\hbar}$	Perturbation theory

All four are the same theory: four lenses on the same abstract vector $|h_t\rangle$ rotating under the same Hamiltonian H .

The Heisenberg formulation is the energy eigenbasis: the state is $c_n = \langle E_n | h_t \rangle$, and each component acquires a phase $e^{-iE_n t}$. The dynamics is diagonal, trivial, and algebraic—but reading off physical predictions (position probabilities, scattering cross sections) requires changing basis.

The Schrödinger formulation is the position eigenbasis: the state is $\psi(x, t) = \langle x | h_t \rangle$, and the equation is a PDE. The dynamics is a differential operator—harder than a diagonal matrix, but natural for single-particle problems where the potential $V(x)$ is explicit.

The path integral is the position eigenbasis at many time slices: the propagator $\langle x_f | e^{-iHT} | x_i \rangle$ is expressed as a sum over intermediate positions, with each path weighted by $e^{iS[x]/\hbar}$. It is the matrix-power formula of Section 1 in the position basis, and it is the most powerful tool for perturbative calculations (Feynman diagrams) and non-perturbative techniques (lattice Monte Carlo).

The Dirac formulation commits to no basis. The vector $|h_t\rangle$ exists independently; the equation $\dot{h}_t = -iHh_t$ is a matrix-vector ODE; and any basis can be chosen at any time by inserting the appropriate resolution of the identity $\mathbf{1} = \sum_\lambda |\lambda\rangle \langle \lambda|$.

Why the Dirac formulation wins for QFT. For quantum field theory, the position-basis formulations (Schrödinger and path integral) become cumbersome: the “position” is now a field configuration $\phi(\mathbf{x})$, the integral $\int \mathcal{D}\phi$ is over an infinite-dimensional function space, and the measure $\mathcal{D}\phi$ requires careful definition (regularisation, renormalisation, Wick rotation). These difficulties are real but technical—they are difficulties of the *basis choice*, not of the physics.

The Dirac formulation avoids them entirely. The Fock vector $|h_t\rangle \in \mathcal{F}$ is a single object, encoding all particle numbers in a superposition. The Hamiltonian H is a polynomial of creation and annihilation operators. The equation $\dot{h}_t = -iHh_t$ is a first-order ODE for a vector in Fock space, identical in form to a linear RNN update. No path integral is needed. No functional measure

is needed. No position-basis wavefunction is needed. The Fock vector is the hidden state; the Hamiltonian is the weight matrix; the Born rule is the output layer. This is the Fock-Born framework.

The path integral re-enters when we need to compute specific matrix elements of e^{-iHT} —transition amplitudes, correlation functions, scattering amplitudes. It is a *computational tool* for evaluating Fock-layer quantities, not a replacement for the Fock vector. The Lagrangian, the action, and $\int \mathcal{D}\phi e^{iS[\phi]}$ are the scaffold; the Fock vector $|h_t\rangle$ and the Hamiltonian H are the building.

The remainder of this book develops both tools—the Hamiltonian (Fock vector, creation/annihilation operators, Schrödinger RNN) and the Lagrangian (path integral, Feynman diagrams, generating functionals)—and uses whichever is more convenient for each calculation. But the foundational language is always the Dirac formulation: $|h_t\rangle, H, \dot{h}_t = -iHh_t$. This is the equation that works at every level—single particle, many particles, quantum field theory, the Standard Model—without changing its form.

6 The Harmonic Oscillator: A Complete Treatment

The quantum harmonic oscillator is the most important system in physics. It is exactly solvable, its eigenstates form the basis of quantum field theory, and its ladder operators are the creation and annihilation operators that build the Fock space of the universe. Every free field is an infinite collection of harmonic oscillators, one per momentum mode. Understanding the single oscillator completely is understanding the free field completely.

6.1 The Hamiltonian, ladder operators, and the spectrum

The classical harmonic oscillator has Hamiltonian $H_{\text{cl}} = p^2/2m + m\omega^2 x^2/2$. Setting $m = 1$, $\omega = 1$ for clarity (we restore ω when needed):

$$H = \frac{1}{2}(\hat{P}^2 + \hat{X}^2), \quad [\hat{X}, \hat{P}] = i\mathbf{1}. \quad (6.1)$$

In the position basis, $\hat{X} = x \cdot$ and $\hat{P} = -id/dx$, so $H = \frac{1}{2}(-d^2/dx^2 + x^2)$. On a discrete lattice with spacing s , H is a concrete Hermitian matrix: the sum of a tridiagonal kinetic matrix ($\hat{P}^2$) and a diagonal potential matrix ($\hat{X}^2$).

We now solve this system entirely algebraically—no differential equation, no wavefunction, just commutators and inner products. The key idea: factorise H into a product of simpler operators, exactly as one might factorise $x^2 + p^2 = (x - ip)(x + ip)$ for ordinary numbers (except that $\hat{X}$ and $\hat{P}$ do not commute, so the factorisation is not exact and produces a remainder).

Defining the ladder operators

Define two operators built from $\hat{X}$ and $\hat{P}$:

$$a = \frac{1}{\sqrt{2}}(\hat{X} + i\hat{P}), \quad a^\dagger = \frac{1}{\sqrt{2}}(\hat{X} - i\hat{P}). \quad (6.2)$$

The operator a is called the *annihilation* (or lowering) operator; $a^\dagger$ is the *creation* (or raising) operator. The names will be justified shortly.

Deriving the commutator $[a, a^\dagger] = 1$. Compute $[a, a^\dagger]$ directly from the definition:

$$\begin{aligned} [a, a^\dagger] &= \frac{1}{2}[(\hat{X} + i\hat{P}), (\hat{X} - i\hat{P})] \\ &= \frac{1}{2}([\hat{X}, \hat{X}] - i[\hat{X}, \hat{P}] + i[\hat{P}, \hat{X}] - i^2[\hat{P}, \hat{P}]) \\ &= \frac{1}{2}(0 - i \cdot i + i \cdot (-i) - (-1) \cdot 0) \\ &= \frac{1}{2}(1 + 1) = 1. \end{aligned} \quad (6.3)$$

We used $[\hat{X}, \hat{P}] = i$, $[\hat{P}, \hat{X}] = -i$, and $[\hat{X}, \hat{X}] = [\hat{P}, \hat{P}] = 0$. The result:

$$\boxed{[a, a^\dagger] = 1.} \quad (6.4)$$

This single relation is the algebraic engine that determines the entire spectrum.

Expressing H in terms of a and $a^\dagger$. Invert (11.5):

$$\hat{X} = \frac{a + a^\dagger}{\sqrt{2}}, \quad \hat{P} = \frac{a - a^\dagger}{i\sqrt{2}}. \quad (6.5)$$

Substitute into $H = \frac{1}{2}(\hat{P}^2 + \hat{X}^2)$:

$$\hat{X}^2 = \frac{1}{2}(a + a^\dagger)^2 = \frac{1}{2}(a^2 + aa^\dagger + a^\dagger a + a^{\dagger 2}), \quad (6.6)$$

$$\hat{P}^2 = -\frac{1}{2}(a - a^\dagger)^2 = -\frac{1}{2}(a^2 - aa^\dagger - a^\dagger a + a^{\dagger 2}). \quad (6.7)$$

Adding:

$$\hat{X}^2 + \hat{P}^2 = \frac{1}{2}(aa^\dagger + a^\dagger a) + \frac{1}{2}(aa^\dagger + a^\dagger a) = aa^\dagger + a^\dagger a. \quad (6.8)$$

(The a^2 and $a^{\dagger 2}$ terms cancel.) Now use the commutator $[a, a^\dagger] = 1$ to rewrite $aa^\dagger = a^\dagger a + 1$:

$$\hat{X}^2 + \hat{P}^2 = (a^\dagger a + 1) + a^\dagger a = 2a^\dagger a + 1. \quad (6.9)$$

Therefore:

$$H = \frac{1}{2}(\hat{X}^2 + \hat{P}^2) = a^\dagger a + \frac{1}{2} = \hat{n} + \frac{1}{2}, \quad (6.10)$$

where $\hat{n} \equiv a^\dagger a$ is the *number operator*. The $+1/2$ is the zero-point energy: it arises from the non-commutativity $[a, a^\dagger] = 1$, which makes $aa^\dagger \neq a^\dagger a$. If position and momentum commuted (as in classical mechanics), $aa^\dagger = a^\dagger a$ and the zero-point energy would vanish.

The spectrum: algebraic derivation, step by step

We now derive the complete spectrum of H —all eigenvalues and all eigenstates—using only the commutation relation $[a, a^\dagger] = 1$ and the Hermiticity of H . No differential equation is solved; no wavefunction is written. This is the power of the algebraic method, and it generalises directly to angular momentum, the Lorentz group, and the gauge groups of the Standard Model.

Step 1: the commutator of H with a and $a^\dagger$. Compute $[H, a^\dagger]$ using $H = a^\dagger a + 1/2$:

$$[H, a^\dagger] = [a^\dagger a, a^\dagger] = a^\dagger[a, a^\dagger] + [a^\dagger, a^\dagger]a = a^\dagger \cdot 1 + 0 = a^\dagger. \quad (6.11)$$

(We used the identity $[AB, C] = A[B, C] + [A, C]B$ with $A = a^\dagger$, $B = a$, $C = a^\dagger$.) Similarly:

$$[H, a] = [a^\dagger a, a] = a^\dagger[a, a] + [a^\dagger, a]a = 0 + (-1) \cdot a = -a. \quad (6.12)$$

These two relations are the key: $a^\dagger$ commutes with H up to $+a^\dagger$, and a commutes with H up to $-a$.

Step 2: $a^\dagger$ raises the energy by exactly 1. Suppose $|E\rangle$ is an eigenstate of H with eigenvalue E : $H|E\rangle = E|E\rangle$. Apply H to $a^\dagger|E\rangle$:

$$\begin{aligned} H(a^\dagger|E\rangle) &= (Ha^\dagger)|E\rangle = ([H, a^\dagger] + a^\dagger H)|E\rangle \\ &= (a^\dagger + a^\dagger H)|E\rangle = a^\dagger|E\rangle + a^\dagger E|E\rangle = (E+1)(a^\dagger|E\rangle). \end{aligned} \quad (6.13)$$

Therefore: $a^\dagger|E\rangle$ is an eigenstate of H with eigenvalue $E+1$. The creation operator raises the energy by exactly one unit. Applied repeatedly: $a^{\dagger k}|E\rangle$ has energy $E+k$.

Step 3: a lowers the energy by exactly 1. The same argument with $[H, a] = -a$:

$$H(a|E\rangle) = ([H, a] + aH)|E\rangle = (-a + aE)|E\rangle = (E-1)(a|E\rangle). \quad (6.14)$$

The annihilation operator lowers the energy by one unit.

Step 4: the energy is bounded below (the ladder must terminate). For any state $|\psi\rangle$:

$$\begin{aligned}\langle\psi|H|\psi\rangle &= \langle\psi|(a^\dagger a + \tfrac{1}{2})|\psi\rangle = \langle\psi|a^\dagger a|\psi\rangle + \tfrac{1}{2} \\ &= \|a|\psi\rangle\|^2 + \tfrac{1}{2} \geq \tfrac{1}{2}.\end{aligned}\tag{6.15}$$

(We used $\langle\psi|a^\dagger a|\psi\rangle = (a|\psi\rangle)^\dagger(a|\psi\rangle) = \|a|\psi\rangle\|^2 \geq 0$.) The energy expectation value is at least $1/2$ in any state. Therefore the lowering sequence $|E\rangle, a|E\rangle, a^2|E\rangle, \dots$ with energies $E, E-1, E-2, \dots$ cannot continue indefinitely: it must terminate at some ground state $|0\rangle$ satisfying:

$$a|0\rangle = 0.\tag{6.16}$$

(If $a|0\rangle \neq 0$, then $a|0\rangle$ would have energy $E_0 - 1 < E_0$, contradicting the assumption that $|0\rangle$ is the ground state.)

Step 5: the ground state energy is $1/2$. Apply H to $|0\rangle$:

$$H|0\rangle = (a^\dagger a + \tfrac{1}{2})|0\rangle = a^\dagger \underbrace{(a|0\rangle)}_{=0} + \tfrac{1}{2}|0\rangle = \tfrac{1}{2}|0\rangle.\tag{6.17}$$

The ground state energy is $E_0 = 1/2$. This is the zero-point energy: the minimum energy enforced by the uncertainty principle ($\Delta x \Delta p \geq 1/2$ means the particle cannot sit still at $x = 0$).

Step 6: the normalization of $a^\dagger|n\rangle$. The state $a^\dagger|n\rangle$ has energy $E_n + 1 = n + 3/2$, so it is proportional to $|n+1\rangle$. To find the proportionality constant, compute its norm:

$$\begin{aligned}\|a^\dagger|n\rangle\|^2 &= \langle n|aa^\dagger|n\rangle = \langle n|(a^\dagger a + 1)|n\rangle \\ &= \langle n|(\hat{n} + 1)|n\rangle = (n+1)\langle n|n\rangle = n+1.\end{aligned}\tag{6.18}$$

(We used $aa^\dagger = a^\dagger a + [a, a^\dagger] = \hat{n} + 1$ and $\hat{n}|n\rangle = n|n\rangle$.) Since $\|a^\dagger|n\rangle\|^2 = n+1$:

$$\boxed{a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle}.\tag{6.19}$$

Similarly, $\|a|n\rangle\|^2 = \langle n|a^\dagger a|n\rangle = n$, so:

$$\boxed{a|n\rangle = \sqrt{n}|n-1\rangle}.\tag{6.20}$$

These normalizations are not conventions—they are consequences of $[a, a^\dagger] = 1$. The $\sqrt{n+1}$ and $\sqrt{n}$ factors are the reason the concrete matrices in Section 6 have square-root entries on the off-diagonal.

Step 7: the complete spectrum. Starting from $|0\rangle$ and applying $a^\dagger$ repeatedly:

$$|n\rangle = \frac{(a^\dagger)^n}{\sqrt{n!}}|0\rangle, \quad n = 0, 1, 2, \dots\tag{6.21}$$

The $\sqrt{n!}$ in the denominator accumulates from the normalization factors at each step: $a^\dagger|0\rangle = \sqrt{1}|1\rangle$, $(a^\dagger)^2|0\rangle = \sqrt{1} \cdot \sqrt{2}|2\rangle = \sqrt{2!}|2\rangle$, $\dots$, $(a^\dagger)^n|0\rangle = \sqrt{n!}|n\rangle$.

Each $|n\rangle$ is an eigenstate of H with eigenvalue:

$$\boxed{E_n = n + \tfrac{1}{2}, \quad n = 0, 1, 2, \dots}\tag{6.22}$$

The spectrum is equally spaced, starting at $E_0 = 1/2$. Each application of $a^\dagger$ adds one quantum of energy.

Energy in the quantum framework. In the energy eigenbasis, the Schrödinger RNN acts as:

$$h_t = e^{-iHt}h_0 = \sum_{n=0}^{\infty} c_n e^{-i(n+1/2)t} |n\rangle. \quad (6.23)$$

Each Fock state $|n\rangle$ phase-rotates at angular frequency $E_n = n + 1/2$. The energy is the rotation rate of the engine: the $|n\rangle$ component rotates $n + 1/2$ times per 2π time units. The Born probability for finding n quanta is:

$$P(n) = |c_n|^2 \quad (\text{time-independent: the rotation preserves } |c_n|^2). \quad (6.24)$$

Energy eigenstates are the directions in Hilbert space that do not mix under the engine's evolution. They are the principal components of the dynamics.

6.2 The ground state and excited states: explicit wavefunctions

The algebraic method gives the spectrum without wavefunctions. But the position-basis wavefunctions are useful for visualisation and for connecting to classical intuition.

The ground state wavefunction. The condition $a|0\rangle = 0$ becomes, in the position basis:

$$a\psi_0(x) = \frac{1}{\sqrt{2}}(\hat{X} + i\hat{P})\psi_0(x) = \frac{1}{\sqrt{2}}\left(x + \frac{d}{dx}\right)\psi_0(x) = 0. \quad (6.25)$$

(We used $\hat{X}\psi(x) = x\psi(x)$ and $\hat{P}\psi(x) = -i\psi'(x)$, so $i\hat{P}\psi(x) = \psi'(x)$.) This is a first-order ODE:

$$\psi_0'(x) = -x\psi_0(x). \quad (6.26)$$

Separate variables: $d\psi_0/\psi_0 = -x dx$. Integrate both sides: $\ln \psi_0 = -x^2/2 + \text{const}$. Exponentiate:

$$\psi_0(x) = C e^{-x^2/2}, \quad (6.27)$$

where C is a normalization constant.

Fixing the normalization. The normalization condition $\int_{-\infty}^{\infty} |\psi_0(x)|^2 dx = 1$ requires:

$$|C|^2 \int_{-\infty}^{\infty} e^{-x^2} dx = |C|^2 \sqrt{\pi} = 1, \quad (6.28)$$

where we used the Gaussian integral $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$ (derived by squaring the integral, converting to polar coordinates, and evaluating the radial integral). Therefore $C = \pi^{-1/4}$, and:

$$\boxed{\psi_0(x) = \pi^{-1/4} e^{-x^2/2}}. \quad (6.29)$$

A Gaussian centred at $x = 0$. The Born distribution for position: $p(x) = |\psi_0(x)|^2 = \pi^{-1/2} e^{-x^2}$, a Gaussian with $\sigma_x = 1/\sqrt{2}$.

The momentum wavefunction is the Fourier transform: $\tilde{\psi}_0(p) = \langle p|0\rangle = \int dx e^{-ipx} \psi_0(x)/\sqrt{2\pi} = \pi^{-1/4} e^{-p^2/2}$ (the Fourier transform of a Gaussian is a Gaussian). The momentum Born distribution is also Gaussian with $\sigma_p = 1/\sqrt{2}$. The product: $\sigma_x \sigma_p = 1/\sqrt{2} \cdot 1/\sqrt{2} = 1/2$, saturating the uncertainty bound exactly.

The zero-point energy $E_0 = 1/2$ is the minimum energy consistent with $\Delta x \Delta p \geq 1/2$: a particle localised at $x = 0$ with $p = 0$ would have $\Delta x = \Delta p = 0$, violating the uncertainty principle. The ground state is the tightest compromise.

The first excited state $\psi_1(x)$: a worked example. Apply $a^\dagger$ to ψ_0 :

$$\begin{aligned}\psi_1(x) &= a^\dagger \psi_0(x) = \frac{1}{\sqrt{2}} \left(x - \frac{d}{dx} \right) \psi_0(x) \\ &= \frac{1}{\sqrt{2}} \left(x \cdot \pi^{-1/4} e^{-x^2/2} - \pi^{-1/4} (-x) e^{-x^2/2} \right) \\ &= \frac{1}{\sqrt{2}} \cdot 2x \cdot \pi^{-1/4} e^{-x^2/2} = \sqrt{2} \pi^{-1/4} x e^{-x^2/2}.\end{aligned}\quad (6.30)$$

(We used $a^\dagger = (x - d/dx)/\sqrt{2}$ in the position basis, and $d(e^{-x^2/2})/dx = -xe^{-x^2/2}$.) Check: $\psi_1(x)$ has one node (one zero crossing at $x = 0$), as expected for the first excited state. The Hermite polynomial is $H_1(x) = 2x$, and $\psi_1 = (1/\sqrt{2^1 \cdot 1!}) \pi^{-1/4} H_1(x) e^{-x^2/2} = \sqrt{2} \pi^{-1/4} x e^{-x^2/2}$. ✓

The general pattern. The n -th eigenstate is built by applying $(a^\dagger)^n$ to the ground state:

$$|n\rangle = \frac{(a^\dagger)^n}{\sqrt{n!}} |0\rangle, \quad n = 0, 1, 2, \dots \quad (6.31)$$

In the position basis:

$$\psi_n(x) = \langle x|n\rangle = \frac{1}{\sqrt{2^n n!}} \pi^{-1/4} H_n(x) e^{-x^2/2}, \quad (6.32)$$

where $H_n(x)$ is the n -th (physicist's) Hermite polynomial: $H_0(x) = 1$, $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2$, $H_3(x) = 8x^3 - 12x$. The wavefunction $\psi_n(x)$ has n nodes— n sign changes across the x -axis—because $H_n(x)$ is a polynomial of degree n .

The ladder action, restated for reference:

$$\boxed{a |n\rangle = \sqrt{n} |n-1\rangle, \quad a^\dagger |n\rangle = \sqrt{n+1} |n+1\rangle, \quad a |0\rangle = 0.} \quad (6.33)$$

These three equations—derived from $[a, a^\dagger] = 1$ alone—are the complete specification of the creation and annihilation operators that build the Fock space of quantum field theory. Every free field is an infinite collection of these operators, one pair per mode $\mathbf{k}$, with $[a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'}$. The single-oscillator algebra derived here is the atom from which the entire edifice is built.

6.3 Concrete matrix representations

In the energy eigenbasis $\{|0\rangle, |1\rangle, |2\rangle, \dots\}$, every operator becomes an explicit matrix.

The Hamiltonian is diagonal: $H = \text{diag}(\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \dots)$.

The annihilation operator is superdiagonal:

$$a \approx \begin{pmatrix} 0 & \sqrt{1} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{3} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{4} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad a^\dagger = a^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \sqrt{1} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{4} & 0 \end{pmatrix}. \quad (6.34)$$

Position and momentum in the energy basis: $\hat{X} = (a + a^\dagger)/\sqrt{2}$, $\hat{P} = (a - a^\dagger)/i\sqrt{2}$:

$$\hat{X} \approx \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & \sqrt{1} & 0 & 0 \\ \sqrt{1} & 0 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 0 & \sqrt{3} \\ 0 & 0 & \sqrt{3} & 0 \end{pmatrix}, \quad \hat{P} \approx \frac{i}{\sqrt{2}} \begin{pmatrix} 0 & -\sqrt{1} & 0 & 0 \\ \sqrt{1} & 0 & -\sqrt{2} & 0 \\ 0 & \sqrt{2} & 0 & -\sqrt{3} \\ 0 & 0 & \sqrt{3} & 0 \end{pmatrix}. \quad (6.35)$$

Position and momentum are tridiagonal (off-diagonal) in the energy basis—they mix adjacent Fock sectors. This is the uncertainty principle made concrete: the observables $\hat{X}$ and $\hat{P}$ are not simultaneously diagonal in any basis.

The ground state as a vector: $|0\rangle = (1, 0, 0, \dots)^T$. Verification: $a|0\rangle = 0$ follows from the matrix product of the superdiagonal a with e_1 .

6.4 The Hamiltonian motor: free and interacting

The free motor. $H_0 = \omega(\hat{n} + 1/2)$ is a polynomial of degree 2 in $a, a^\dagger$. It is diagonal in the Fock basis: each sector $|n\rangle$ phase-rotates independently at frequency $\omega(n + 1/2)$. The motor does not create or destroy quanta. It is a collection of independent clocks, one per Fock sector. This is a free theory: the quanta persist, unchanging, rotating in phase forever.

Every Hamiltonian in quantum field theory is a polynomial of creation and annihilation operators, typically of degree ≤ 4 . The game engine of the universe is a polynomial motor. The specific polynomial determines the physics; the architectural form is universal.

The interacting motor. Add a nonlinear perturbation:

$$H = \underbrace{\omega(\hat{n} + \frac{1}{2})}_{H_0} + \underbrace{\frac{\lambda}{3!}(a + a^\dagger)^3 + \frac{\mu}{4!}(a + a^\dagger)^4}_{H_{\text{int}}}. \quad (6.36)$$

Expanding the interaction in normal-ordered form: $(a + a^\dagger)^3 = (a^\dagger)^3 + 3(a^\dagger)^2 a + 3a^\dagger a^2 + a^3 + 3a^\dagger + 3a$. Each monomial is an instruction: $(a^\dagger)^3$ creates 3 quanta; $(a^\dagger)^2 a$ destroys 1, creates 2; a^3 destroys 3.

The motor is no longer diagonal in the Fock basis. It has off-diagonal elements connecting different Fock sectors. At each tick, the motor does two things: it phase-rotates each sector (the H_0 part) and transfers amplitude between sectors (the H_{int} part). The hidden state $h_t = \sum_n c_n(t) |n\rangle$ evolves as a coupled system: the polynomial motor mixes the coefficients $c_n(t)$ at every step.

The motor is a polynomial of a and $a^\dagger$. It does not need a spacetime to operate in. It does not need a metric $g_{\mu\nu}$. It needs only the Fock space, the ladder operators, and the polynomial H .

6.5 The vacuum: one quiet vector

The free vacuum. The ground state of the free Hamiltonian $H_0 = \omega(\hat{n} + 1/2)$ is the state $|0\rangle$ satisfying $a|0\rangle = 0$. In the Fock basis, it is the vector $(1, 0, 0, \dots)^T$: no quanta, all amplitude in the zeroth sector. Its energy is $E_0 = \omega/2$ (the zero-point energy). Under time evolution: $e^{-iH_0 t} |0\rangle = e^{-i\omega t/2} |0\rangle$ —a global phase, nothing more. The free vacuum sits still.

The interacting vacuum. The ground state of the full Hamiltonian $H = H_0 + \lambda V$ is the *interacting vacuum* $|\Omega\rangle$:

$$H |\Omega\rangle = E_\Omega |\Omega\rangle. \quad (6.37)$$

It is a single, definite vector in Hilbert space—the lowest eigenstate of the full Hamiltonian. Under time evolution: $e^{-iHt} |\Omega\rangle = e^{-iE_\Omega t} |\Omega\rangle$ —again a global phase and nothing else. The interacting vacuum, like the free vacuum, *sits still*.

Expanding $|\Omega\rangle$ in the free basis. The free Fock states $\{|n\rangle_0\}$ and the interacting eigenstates $\{|n\rangle\}$ are two different bases of the same Hilbert space. Expanding $|\Omega\rangle$ in the free basis:

$$|\Omega\rangle = c_0 |0\rangle_0 + c_1 |1\rangle_0 + c_2 |2\rangle_0 + c_3 |3\rangle_0 + \cdots \quad (6.38)$$

For a parity-symmetric interaction ($V(\hat{x}) = V(-\hat{x})$, so V commutes with parity $(-1)^{\hat{n}}$), the vacuum has even parity: only even n contribute, and:

$$|\Omega\rangle = c_0 |0\rangle_0 + c_2 |2\rangle_0 + c_4 |4\rangle_0 + \cdots \quad (6.39)$$

The coefficients c_n are computed perturbatively using (6.48):

$$c_n = \langle n_0 | \Omega \rangle \approx \frac{\langle n_0 | V | 0 \rangle_0}{E_0^{(0)} - E_n^{(0)}} + O(\lambda^2), \quad n \neq 0. \quad (6.40)$$

For the quartic interaction $V = \hat{x}^4/4! = (a + a^\dagger)^4/(4! \cdot 4)$, the nonzero matrix elements are $\langle 2 |_0 V | 0 \rangle_0$ and $\langle 4 |_0 V | 0 \rangle_0$ (the terms in $(a + a^\dagger)^4$ that create 2 or 4 quanta from the free vacuum). The coefficient $c_0 \approx 1 - O(\lambda^2)$ (normalisation correction), and $|c_n| \sim \lambda$ for $n \geq 2$: the multi-particle components are small corrections, suppressed by the coupling.

A basis-change analogy for ML readers. Consider a vector $\mathbf{v} = (1, 0, 0)^T$ in 3D (all amplitude on the first axis). Rotate the coordinate system by a small angle θ : in the rotated coordinates, $\mathbf{v} \approx (\cos \theta, -\sin \theta, 0) \approx (1, -\theta, 0)$. In the new coordinates, $\mathbf{v}$ has a small component along the second axis—but the vector has not changed. It is still one vector, pointing in the same direction. The “component along the second axis” is an artefact of the coordinate change, not a property of the vector.

The interacting vacuum $|\Omega\rangle$ is exactly this vector. In its own eigenbasis (the interacting basis), it is $(1, 0, 0, \dots)^T$ —all amplitude on one state, perfectly still. In the free basis, it has small components on $|2\rangle_0$, $|4\rangle_0$, etc.—but these are artefacts of the basis change, not physical events. Nothing has been created. Nothing is being destroyed. The vacuum is one vector, and it is quiet.

Refuting “the vacuum teems with virtual particles.”

The standard language of many textbooks and popular accounts says: “the vacuum is not empty; it teems with virtual particles that pop in and out of existence, borrowing energy from the vacuum and returning it within the time allowed by the uncertainty principle $\Delta E \cdot \Delta t \sim \hbar$.”

Every part of this narrative is a conflation—Born-layer language smuggled into the Fock layer:

“Teems with virtual particles.” The coefficients $c_2, c_4, \dots$ in (6.39) are basis-decomposition coefficients—the coordinates of one quiet vector in a particular basis. They are not particles. No observer detects them. No Born rendering produces them. Calling them “virtual particles” narrates Fock-layer inner products as if they were Born-layer events.

“Pop in and out of existence.” The vacuum $|\Omega\rangle$ is a time-independent eigenstate of H : $e^{-iHt} |\Omega\rangle = e^{-iE_\Omega t} |\Omega\rangle$. Nothing pops in or out. The state does not change. The coefficients c_n are time-independent (because the expansion is in the free basis, which is not the eigenbasis of H , the c_n ’s acquire phases—but $|c_n|^2$ is constant).

“Borrowing energy.” The vacuum has definite energy E_Ω . There is no energy fluctuation. The $\Delta E \cdot \Delta t$ argument confuses two different things: the Born variance of the energy observable (which is zero for an energy eigenstate: $\Delta E = 0$) and the time-dependence of Born distributions for non-energy observables (which can vary, but this is not “energy borrowing”).

“Within the time allowed by the uncertainty principle.” The energy-time uncertainty relation $\Delta E \cdot \Delta t \geq \hbar/2$ is not about “borrowing energy for a time Δt .” It is about the time Δt over

which the Born distribution of a non-energy observable changes appreciably: $\Delta t \sim \hbar/(d\langle\hat{A}\rangle/dt)$ for observable $\hat{A}$. For an energy eigenstate, no Born distribution changes, so $\Delta t = \infty$: the state is stationary.

The correct picture. The interacting vacuum $|\Omega\rangle$ differs from the free vacuum $|0\rangle_0$ because the full Hamiltonian $H = H_0 + \lambda V$ has different eigenstates than H_0 . The difference is real: the vacuum energy $E_\Omega \neq E_0^{(0)}$, the vacuum expectation values $\langle\Omega|\hat{x}^2|\Omega\rangle \neq \langle 0|\hat{x}^2|0\rangle$, and the scattering amplitudes computed from $|\Omega\rangle$ differ from those computed from $|0\rangle_0$. These differences are physical and measurable.

But the mechanism is not “teeming with virtual particles.” The mechanism is that the ground state of H is a different vector from the ground state of H_0 , in the same Hilbert space. The two vectors are related by a (small) rotation in Hilbert space, and the “virtual-particle content” is the projection of this rotation onto the free-field basis—a basis-decomposition artefact, not a physical process.

In the Fock-Born framework: the interacting vacuum is one quiet vector. It sits still. Its properties (energy, expectation values, correlators) differ from the free vacuum because the motor H that determines the ground state is different. These properties are Fock-layer data. The word “particle” belongs to the Born layer. Applying it to Fock-layer basis-decomposition coefficients is the conflation.

6.6 The catalog of observables: what the Born layer can measure

The Fock-Born framework requires a sharp distinction between Born-layer observables (Hermitian operators submitted to the interface, producing eigenvalues the observer records) and Fock-layer intermediaries (matrix elements and calculational quantities computed inside the engine, never submitted to the Born interface). The harmonic oscillator is the place to make this distinction concrete, because it extends directly to quantum field theory.

Born-layer observables of the single oscillator. These are the Hermitian operators the observer can submit to the Born interface, together with their vocabularies and eigenstates:

1. *Number operator* $\hat{n} = a^\dagger a$. Vocabulary: $\{0, 1, 2, \dots\}$ (non-negative integers). Eigenstates: $|n\rangle$ (the Fock basis). Born probability: $P(n) = |c_n|^2$. This is the observable that counts quanta. For the free oscillator, $[\hat{n}, H_0] = 0$: the number of quanta is conserved. For the interacting oscillator, $[\hat{n}, H_{\text{int}}] \neq 0$: the Born distribution over n changes with time.

2. *Position* $\hat{x} = (a + a^\dagger)/\sqrt{2\omega}$. Vocabulary: $\{x \in \mathbb{R}\}$. Eigenstates: $|x\rangle$ (position eigenstates, delta functions). Born probability: $p(x) = |\psi(x)|^2$. This is the “field value” observable. In QFT, $\hat{\phi}(\mathbf{x})$ is the multi-mode generalisation.

3. *Momentum* $\hat{p} = -i\sqrt{\omega/2}(a - a^\dagger)$. Vocabulary: $\{p \in \mathbb{R}\}$. Eigenstates: $|p\rangle$ (plane waves). Born probability: $p(p) = |\tilde{\psi}(p)|^2$ (squared Fourier transform). Incompatible with $\hat{x}$: $[\hat{x}, \hat{p}] = i$, so $\Delta x \Delta p \geq 1/2$.

4. *Energy* H . Vocabulary: $\{E_n = \omega(n + 1/2)\}$ for the free oscillator, or $\{E_n\}$ (the eigenvalues of the full H) for the interacting case. Eigenstates: $|E_n\rangle$. For the free oscillator, $H = \omega(\hat{n} + 1/2)$ shares its eigenbasis with $\hat{n}$; for the interacting oscillator, the energy eigenstates differ from the Fock states and must be computed (e.g. by diagonalising the truncated matrix H_{mn}).

5. *Quadrature operators* $\hat{X}_\theta = (ae^{-i\theta} + a^\dagger e^{i\theta})/\sqrt{2}$. A continuous family parameterised by angle θ . At $\theta = 0$, this is $\hat{x}$; at $\theta = \pi/2$, it is $\hat{p}$ (up to normalisation). In quantum optics, homodyne detection measures $\hat{X}_\theta$ by mixing the signal with a local oscillator at phase θ . This is the observable of choice in gravitational-wave detectors (LIGO).

6. *Parity* $(-1)^{\hat{n}}$. Vocabulary: $\{+1, -1\}$ (even or odd quanta). Eigenstates: $|n\rangle$ with eigenvalue $(-1)^n$. This is a highly degenerate observable: eigenvalue $+1$ has eigensubspace spanned by all even $|n\rangle$, eigenvalue -1 by all odd $|n\rangle$.

Born-layer observables in QFT (multi-mode). When the oscillator is promoted to a field (one oscillator per mode $\mathbf{k}$), the observable catalog expands:

7. *Total particle number* $\hat{N} = \sum_k a_k^\dagger a_k$. Counts quanta across all modes. Conserved by the free Hamiltonian; broken by interactions.

8. *Field at a point* $\hat{\phi}(\mathbf{x})$. A superposition of all mode operators. Its Born distribution in the vacuum is Gaussian with zero mean and variance $\langle 0 | \hat{\phi}^2(\mathbf{x}) | 0 \rangle \neq 0$. This nonzero variance is what the phrase “vacuum fluctuations” actually refers to: it is a property of the Born distribution, not a temporal process in the Fock layer.

9. *Conserved charges*. For complex fields: charge $\hat{Q} = \hat{N}_{\text{particle}} - \hat{N}_{\text{antiparticle}}$, which is conserved even by interactions ($[\hat{Q}, H] = 0$ for $U(1)$ -invariant Hamiltonians). For gauge theories: colour charge, isospin, baryon number, lepton number. Conserved charges are Born-layer observables whose Born distributions are exactly delta (definite eigenvalue) if the state is prepared in a charge sector.

10. *Macroscopic observables*. Magnetisation $\hat{M} = \sum_i \hat{S}_i^z$, order parameters, collective field averages $\hat{\Phi} = V^{-1} \int d^3x \hat{\phi}(\mathbf{x})$. These have Born distributions that are near-delta for large systems ($\text{Var} \sim 1/N$): the classical regime of subsection 5.1.

What experimentalists actually report. In the interacting theory, the Born-layer observables that appear in experimental papers are:

Cross sections σ : event rates at detectors, proportional to $|\langle f | S | i \rangle|^2$. These are Born probabilities for the scattering observable—the S -matrix projected onto specific final states.

Masses: eigenvalues of the full interacting Hamiltonian H , or equivalently poles of the propagator. The physical mass is a Born-layer quantity: it is what a spectrometer reports.

Lifetimes and decay widths Γ : inverse lifetimes of unstable states, extracted from the Born distribution’s time dependence (the Weisskopf–Wigner exponential).

Spectral lines: energy differences $E_m - E_n$, directly measurable by spectroscopy.

Conserved quantum numbers: eigenvalues of charges $[\hat{Q}, H] = 0$, always definite when the state is in a charge sector.

Fock-layer intermediaries: not observables. The following quantities are computed inside the Fock layer as steps toward Born-layer predictions. They are matrix elements and calculational tools. No observer submits them to the Born interface:

Propagators $\langle \Omega | T \{ \hat{\phi}(x) \hat{\phi}(y) \} | \Omega \rangle$: Fock-layer matrix elements of time-ordered field products. They encode the correlation structure of the vacuum but are not eigenvalues of any Hermitian operator.

Vertex functions and self-energies: amputated, one-particle-irreducible subdiagrams. They are bookkeeping devices in the perturbative expansion.

Running couplings $g(\mu)$: the μ -dependence of the coupling is a Fock-layer re-parameterisation. What the observer measures is a cross section at a specific energy, not the coupling itself.

Wave-function renormalisation Z : a normalisation factor that cancels in physical predictions.

The LSZ reduction formula is the bridge between the two columns: it converts Fock-layer correlation functions (propagators, vertex functions) into Born-layer scattering amplitudes (S -matrix elements). The Feynman diagram machinery operates entirely in the left column—Fock-layer intermediaries—and the LSZ formula is the exit gate to the right column—Born-layer observables.

The conflation catalogued in the preface—“virtual particles,” “the vacuum teems”—is precisely the error of narrating the left column in the language of the right column: treating Fock-layer intermediaries as if they were Born-layer events.

6.7 Perturbation theory and the Dyson series

When H_{int} is small compared to H_0 , the eigenvalues and eigenstates of the full $H = H_0 + H_{\text{int}}$ can be computed order by order as a power series in the coupling. This is perturbation theory: the most widely used tool in quantum mechanics and quantum field theory. Every Feynman diagram computation is a perturbative expansion. We derive it here in full, because the single-oscillator case contains all the structure of the QFT case.

Time-independent perturbation theory: energy corrections

Write $H = H_0 + \lambda V$, where H_0 has known eigenstates $|n^{(0)}\rangle$ with eigenvalues $E_n^{(0)}$, and λV is a small perturbation ($\lambda \ll 1$). Expand the full eigenvalue and eigenstate as power series in λ :

$$E_n = E_n^{(0)} + \lambda E_n^{(1)} + \lambda^2 E_n^{(2)} + \dots, \quad (6.41)$$

$$|n\rangle = |n^{(0)}\rangle + \lambda |n^{(1)}\rangle + \lambda^2 |n^{(2)}\rangle + \dots \quad (6.42)$$

Substitute into $H|n\rangle = E_n|n\rangle$:

$$(H_0 + \lambda V)(|n^{(0)}\rangle + \lambda |n^{(1)}\rangle + \dots) = (E_n^{(0)} + \lambda E_n^{(1)} + \dots)(|n^{(0)}\rangle + \lambda |n^{(1)}\rangle + \dots). \quad (6.43)$$

Expand and collect powers of λ :

Order λ^0 : $H_0|n^{(0)}\rangle = E_n^{(0)}|n^{(0)}\rangle$. ✓ (the unperturbed eigenvalue equation).

Order λ^1 :

$$H_0|n^{(1)}\rangle + V|n^{(0)}\rangle = E_n^{(0)}|n^{(1)}\rangle + E_n^{(1)}|n^{(0)}\rangle. \quad (6.44)$$

First-order energy correction. Project (6.44) onto $\langle n^{(0)}|$:

$$\underbrace{\langle n^{(0)}|H_0|n^{(1)}\rangle}_{E_n^{(0)}\langle n^{(0)}|n^{(1)}\rangle} + \langle n^{(0)}|V|n^{(0)}\rangle = E_n^{(0)}\langle n^{(0)}|n^{(1)}\rangle + E_n^{(1)}\underbrace{\langle n^{(0)}|n^{(0)}\rangle}_{=1}. \quad (6.45)$$

The first terms on each side cancel, leaving:

$$\boxed{E_n^{(1)} = \langle n^{(0)}|V|n^{(0)}\rangle = V_{nn}.} \quad (6.46)$$

The first-order energy correction is the diagonal matrix element of the perturbation in the unperturbed basis: the “self-energy” of state $|n^{(0)}\rangle$ under the interaction V . It is a single number computed from the Fock-layer matrix V_{nn} —not a Born-layer event.

First-order state correction. Project (6.44) onto $\langle m^{(0)}|$ with $m \neq n$:

$$E_m^{(0)}\langle m^{(0)}|n^{(1)}\rangle + V_{mn} = E_n^{(0)}\langle m^{(0)}|n^{(1)}\rangle, \quad (6.47)$$

so:

$$\langle m^{(0)} | n^{(1)} \rangle = \frac{V_{mn}}{E_n^{(0)} - E_m^{(0)}}, \quad m \neq n. \quad (6.48)$$

The corrected state:

$$|n\rangle = |n^{(0)}\rangle + \lambda \sum_{m \neq n} \frac{V_{mn}}{E_n^{(0)} - E_m^{(0)}} |m^{(0)}\rangle + O(\lambda^2). \quad (6.49)$$

The perturbation mixes $|n^{(0)}\rangle$ with other unperturbed states $|m^{(0)}\rangle$, with amplitude proportional to the matrix element V_{mn} divided by the energy difference. States that are close in energy ($E_n^{(0)} \approx E_m^{(0)}$) mix strongly; states that are far apart mix weakly. This is the origin of the “resonance” concept: perturbation theory breaks down when two levels become degenerate ($E_n^{(0)} = E_m^{(0)}$), because the denominator diverges.

Second-order energy correction. Collecting terms at order λ^2 and projecting onto $\langle n^{(0)} |$:

$$E_n^{(2)} = \sum_{m \neq n} \frac{|V_{mn}|^2}{E_n^{(0)} - E_m^{(0)}} = \sum_{m \neq n} \frac{|\langle m^{(0)} | V | n^{(0)} \rangle|^2}{E_n^{(0)} - E_m^{(0)}}. \quad (6.50)$$

The full energy to second order:

$$E_n = E_n^{(0)} + \lambda V_{nn} + \lambda^2 \sum_{m \neq n} \frac{|V_{mn}|^2}{E_n^{(0)} - E_m^{(0)}} + O(\lambda^3). \quad (6.51)$$

The derivation of (6.50): substitute the first-order state correction (6.49) into the order- λ^2 eigenvalue equation and project onto $\langle n^{(0)} |$. The key step: the term $\langle n^{(0)} | V | n^{(1)} \rangle$ produces $\sum_{m \neq n} V_{nm} \cdot V_{mn} / (E_n^{(0)} - E_m^{(0)}) = \sum_{m \neq n} |V_{mn}|^2 / (E_n^{(0)} - E_m^{(0)})$.

Physical content of the second-order correction. The second-order correction involves a sum over *intermediate states* $m \neq n$. The numerator $|V_{mn}|^2$ measures how strongly the perturbation connects states n and m ; the denominator $E_n^{(0)} - E_m^{(0)}$ penalises large energy differences. Each term in the sum represents a “virtual transition” from n to m and back: the perturbation takes the system from $|n\rangle$ to $|m\rangle$ via one application of V , and back from $|m\rangle$ to $|n\rangle$ via another.

Fock-Born diagnostic. This “virtual transition” is a Fock-layer matrix computation—a product of two matrix elements $V_{nm}V_{mn} = |V_{mn}|^2$ divided by an energy difference. It is *not* a Born-layer event: no observer measures the intermediate state m ; no Born sampling occurs; no Bohr trimming resets the state. The system does not “visit” $|m\rangle$. The second-order correction is a term in a matrix perturbation series—a coefficient in the Taylor expansion of the eigenvalue of $H_0 + \lambda V$ around $\lambda = 0$ —not a narrative of events.

Foreshadowing QFT. In quantum field theory, this structure becomes the Feynman diagram:

Single oscillator	QFT (ϕ^4 theory)
Matrix element V_{mn}	Vertex factor $-i\lambda$
Sum over intermediate states $\sum_m$	Loop integral $\int d^4k / (2\pi)^4$
Energy denominator $1/(E_n - E_m)$	Propagator $i/(k^2 - m^2 + i\varepsilon)$
$ V_{mn} ^2 / (E_n - E_m)$	One-loop self-energy diagram
Second-order energy correction	Mass renormalisation

The single-oscillator perturbation theory is the prototype of all Feynman diagram calculations.

The interaction picture and the Dyson series

Perturbation theory computes the eigenvalues and eigenstates of H . The Dyson series computes the time-evolution operator $U(t) = e^{-iHt}$ as a power series. It is the time-dependent version of perturbation theory, and it is the starting point for scattering theory.

The interaction picture. In the Schrödinger picture, the full evolution is $|\psi_S(t)\rangle = e^{-i(H_0+\lambda V)t} |\psi(0)\rangle$. Define the *interaction-picture* state by factoring out the free evolution:

$$|\psi_I(t)\rangle = e^{iH_0t} |\psi_S(t)\rangle. \quad (6.52)$$

This strips off the “trivial” part of the evolution (the phase-rotation by H_0) and keeps only the “interesting” part (the effect of λV). Taking the time derivative:

$$\begin{aligned} i \frac{d}{dt} |\psi_I(t)\rangle &= i \frac{d}{dt} (e^{iH_0t} |\psi_S(t)\rangle) \\ &= -H_0 e^{iH_0t} |\psi_S(t)\rangle + e^{iH_0t} \underbrace{i \frac{d}{dt} |\psi_S(t)\rangle}_{(H_0+\lambda V)|\psi_S(t)\rangle} \\ &= e^{iH_0t} \lambda V |\psi_S(t)\rangle = \underbrace{e^{iH_0t} \lambda V e^{-iH_0t}}_{\lambda V_I(t)} |\psi_I(t)\rangle. \end{aligned} \quad (6.53)$$

The interaction-picture equation of motion is:

$$\boxed{i \frac{d}{dt} |\psi_I(t)\rangle = \lambda V_I(t) |\psi_I(t)\rangle, \quad V_I(t) = e^{iH_0t} V e^{-iH_0t}.} \quad (6.54)$$

The state evolves under the interaction alone, but the interaction $V_I(t) = e^{iH_0t} V e^{-iH_0t}$ now has explicit time dependence inherited from the free evolution.

Iterating the integral equation. Equation (6.54) is equivalent to the integral equation:

$$|\psi_I(t)\rangle = |\psi_I(0)\rangle - i\lambda \int_0^t dt_1 V_I(t_1) |\psi_I(t_1)\rangle. \quad (6.55)$$

Iterate: substitute the right side into itself, once, twice, three times:

Zeroth order: $|\psi_I(t)\rangle \approx |\psi_I(0)\rangle$ (no interaction).

First order: substitute the zeroth-order result into the integral:

$$|\psi_I(t)\rangle \approx |\psi_I(0)\rangle - i\lambda \int_0^t dt_1 V_I(t_1) |\psi_I(0)\rangle. \quad (6.56)$$

Second order: substitute the first-order result:

$$|\psi_I(t)\rangle \approx |\psi_I(0)\rangle - i\lambda \int_0^t dt_1 V_I(t_1) |\psi_I(0)\rangle + (-i\lambda)^2 \int_0^t dt_1 \int_0^{t_1} dt_2 V_I(t_1) V_I(t_2) |\psi_I(0)\rangle. \quad (6.57)$$

Note the nested integration limits: $0 \leq t_2 \leq t_1 \leq t$. This ensures that the operators appear in *time order*: $V_I(t_1)$ acts after $V_I(t_2)$ (because $t_1 \geq t_2$). This ordering matters because $V_I(t_1)$ and $V_I(t_2)$ do not commute at different times.

The Dyson series. Continuing to all orders:

$$|\psi_I(t)\rangle = U_I(t) |\psi_I(0)\rangle, \quad U_I(t) = \sum_{n=0}^{\infty} (-i\lambda)^n \int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n V_I(t_1) V_I(t_2) \cdots V_I(t_n). \quad (6.58)$$

This is the Dyson series: the time-evolution operator in the interaction picture, expanded as a power series in the coupling λ . The time-ordered integration enforces causality: later interactions act to the left. The compact notation:

$$U_I(t) = T \exp\left(-i\lambda \int_0^t V_I(t') dt'\right), \quad (6.59)$$

where T is the time-ordering symbol that places later times to the left.

What each term is. The n -th term in the Dyson series is:

$$U_I^{(n)}(t) = (-i\lambda)^n \int_0^t dt_1 \cdots \int_0^{t_{n-1}} dt_n V_I(t_1) \cdots V_I(t_n). \quad (6.60)$$

It represents n applications of the interaction V , at times $t_n \leq \cdots \leq t_1$, with free evolution $e^{-iH_0(t_k - t_{k+1})}$ between each pair. In the harmonic oscillator, V is a polynomial of a and $a^\dagger$, so each application creates and/or destroys quanta. The matrix element $\langle f | U_I^{(n)}(t) | i \rangle$ is the amplitude for the state to evolve from $|i\rangle$ to $|f\rangle$ via n interaction events— n applications of the polynomial V .

Fock-Born diagnostic. Each term is a Fock-layer matrix computation: a product of matrix elements of V , integrated over intermediate times, summed over intermediate states. The “interaction events at $t_1, \dots, t_n$ ” are integration variables in a perturbative expansion of the matrix exponential e^{-iHt} . They are not Born-layer detections. No observer measures anything at t_k . The system does not “visit” an intermediate state. The Dyson series is a Taylor expansion of a matrix function—algebraic bookkeeping, not a physical narrative.

The QFT generalisation. In quantum field theory, V is replaced by $H_{\text{int}} = \frac{\lambda}{4!} \int d^3x : \hat{\phi}^4 :$, and each application of H_{int} creates and destroys particles across all momentum modes simultaneously. The Dyson series becomes the sum over Feynman diagrams: each diagram is one term, the vertices are the applications of H_{int} , the internal lines are the free propagators $e^{-iH_0(t_k - t_{k+1})}$ connecting successive vertices, and the loop integrals $\int d^4k / (2\pi)^4$ are the sums over intermediate momenta. The single-oscillator Dyson series derived here is the one-mode prototype of this machinery.

6.8 Coherent states and the classical limit

The Fock states $|n\rangle$ are energy eigenstates: definite energy, indefinite position, stationary up to a phase. Nothing in the Fock basis resembles a classical particle oscillating back and forth.

The **coherent state** is the Fock-layer state whose Born-layer rendering is a classical trajectory.

Definition and derivation of the Fock expansion

The coherent state $|\alpha\rangle$, labelled by $\alpha \in \mathbb{C}$, is defined as the eigenstate of a :

$$a |\alpha\rangle = \alpha |\alpha\rangle. \quad (6.61)$$

Note that a is not Hermitian, so its eigenvalue α can be complex, and its eigenstates are not orthogonal.

Derivation of the Fock-basis expansion. Expand $|\alpha\rangle$ in the Fock basis: $|\alpha\rangle = \sum_{n=0}^{\infty} c_n |n\rangle$. Apply a to both sides:

$$a|\alpha\rangle = \sum_{n=0}^{\infty} c_n a|n\rangle = \sum_{n=1}^{\infty} c_n \sqrt{n} |n-1\rangle = \sum_{n=0}^{\infty} c_{n+1} \sqrt{n+1} |n\rangle, \quad (6.62)$$

where we used $a|n\rangle = \sqrt{n}|n-1\rangle$ and shifted the summation index $n \rightarrow n+1$ in the last step. The right side of $a|\alpha\rangle = \alpha|\alpha\rangle$ is:

$$\alpha|\alpha\rangle = \sum_{n=0}^{\infty} \alpha c_n |n\rangle. \quad (6.63)$$

Matching coefficients of $|n\rangle$ on both sides:

$$c_{n+1} \sqrt{n+1} = \alpha c_n. \quad (6.64)$$

This is a recurrence relation. Solve it:

$$c_{n+1} = \frac{\alpha}{\sqrt{n+1}} c_n \Rightarrow c_n = \frac{\alpha^n}{\sqrt{n!}} c_0, \quad (6.65)$$

obtained by iterating: $c_1 = \alpha c_0 / \sqrt{1!}$, $c_2 = \alpha c_1 / \sqrt{2} = \alpha^2 c_0 / \sqrt{2!}$, and so on.

Fixing the normalisation. The normalisation $\langle\alpha|\alpha\rangle = 1$ requires:

$$1 = \sum_{n=0}^{\infty} |c_n|^2 = |c_0|^2 \sum_{n=0}^{\infty} \frac{|\alpha|^{2n}}{n!} = |c_0|^2 e^{|\alpha|^2}. \quad (6.66)$$

(We recognised the Taylor series $\sum x^n/n! = e^x$.) Therefore $c_0 = e^{-|\alpha|^2/2}$ (choosing the conventional real positive phase), and:

$$|\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle. \quad (6.67)$$

Every step was algebra: expand in the Fock basis, apply a , match coefficients, solve the recurrence, normalise. The coherent state is not a postulate; it is the unique solution of $a|\alpha\rangle = \alpha|\alpha\rangle$ in the Fock basis.

The Born distribution over quanta: Poisson. The probability of finding n quanta:

$$P(n) = |\langle n|\alpha\rangle|^2 = e^{-|\alpha|^2} \frac{|\alpha|^{2n}}{n!}, \quad (6.68)$$

which is a Poisson distribution with mean $\bar{n} = |\alpha|^2$ and standard deviation $\Delta n = |\alpha|$. The relative fluctuation:

$$\frac{\Delta n}{\bar{n}} = \frac{1}{|\alpha|} \xrightarrow{|\alpha| \rightarrow \infty} 0. \quad (6.69)$$

For $|\alpha| \gg 1$: the number Born distribution becomes near-delta—the quantum number is effectively definite. This is the classical regime for particle number.

Classical trajectory from the Born layer. The Born expected values of position and momentum in the coherent state (restoring ω):

$$\langle\alpha|\hat{x}|\alpha\rangle = \sqrt{\frac{2}{\omega}} \operatorname{Re}(\alpha e^{-i\omega t}), \quad \langle\alpha|\hat{p}|\alpha\rangle = -\sqrt{2\omega} \operatorname{Im}(\alpha e^{-i\omega t}), \quad (6.70)$$

trace out a classical ellipse in phase space—exactly the trajectory of a classical harmonic oscillator with amplitude $|\alpha|$ and phase $\arg \alpha$.

To derive the first: $\hat{x} = (a + a^\dagger)/\sqrt{2\omega}$, so $\langle \hat{x} \rangle = (\langle a \rangle + \langle a^\dagger \rangle)/\sqrt{2\omega}$. Since $a|\alpha\rangle = \alpha|\alpha\rangle$: $\langle a \rangle = \alpha$. Under free evolution: $|\alpha(t)\rangle = |\alpha e^{-i\omega t}\rangle$ (the coherent state remains coherent with a rotating amplitude), so $\langle a \rangle = \alpha e^{-i\omega t}$ and $\langle a^\dagger \rangle = \alpha^* e^{i\omega t}$. Sum: $\langle \hat{x} \rangle = (\alpha e^{-i\omega t} + \alpha^* e^{i\omega t})/\sqrt{2\omega} = \sqrt{2/\omega} \operatorname{Re}(\alpha e^{-i\omega t})$.

The Born *variance* of position is $\Delta x = 1/\sqrt{2\omega}$, independent of α : the wave packet maintains its shape and oscillates without spreading.

For $|\alpha| \gg 1$, the Born distribution for position is a narrow Gaussian sliding back and forth along the classical trajectory: the Born rendering of the coherent state looks classical. The Born distribution is near-delta at every instant, so sampling is trivial and trimming inert. This is the classical limit: not a new theory, but the Fock-Born architecture in the regime where the Born distribution has negligible variance.

The displacement operator. $|\alpha\rangle = D(\alpha)|0\rangle$ with $D(\alpha) = e^{\alpha a^\dagger - \alpha^* a}$: the coherent state is the vacuum displaced in phase space. To verify: use the Baker–Campbell–Hausdorff identity for two operators with a c -number commutator, $e^{A+B} = e^A e^B e^{-[A,B]/2}$ when $[A, [A, B]] = [B, [A, B]] = 0$. With $A = \alpha a^\dagger$, $B = -\alpha^* a$: $[A, B] = -|\alpha|^2 [a^\dagger, a] = |\alpha|^2$. Therefore: $D(\alpha) = e^{-|\alpha|^2/2} e^{\alpha a^\dagger} e^{-\alpha^* a}$. Acting on $|0\rangle$: $e^{-\alpha^* a}|0\rangle = |0\rangle$ (since $a|0\rangle = 0$, and the Taylor expansion of $e^{-\alpha^* a}$ applied to $|0\rangle$ gives $|0\rangle$). Then $e^{\alpha a^\dagger}|0\rangle = \sum_n \alpha^n (a^\dagger)^n |0\rangle / n! = \sum_n \alpha^n \sqrt{n!} |n\rangle / n! = \sum_n \alpha^n |n\rangle / \sqrt{n!}$. Multiplying by $e^{-|\alpha|^2/2}$ recovers (6.67). ✓

The coherent state is the vacuum with a condensate: the Born expectation value $\langle \alpha | \hat{x} | \alpha \rangle \neq 0$ is a macroscopic field value carried by the Fock state. In QFT, the multi-mode coherent state $\bigotimes_{\mathbf{k}} |\alpha_{\mathbf{k}}\rangle$ is the Fock-layer state that renders as a classical electromagnetic wave.

6.9 Condensates and spontaneous symmetry breaking

The displaced vacuum is the oscillator analogue of the Higgs mechanism. Consider the anharmonic potential:

$$V(\hat{x}) = -\frac{\mu^2}{2} \hat{x}^2 + \frac{\lambda}{4!} \hat{x}^4, \quad \mu^2 > 0, \lambda > 0. \quad (6.71)$$

This is symmetric under $\hat{x} \rightarrow -\hat{x}$, but its classical minima are at $x_0 = \pm v$ with $v = \sqrt{6\mu^2/\lambda}$, not at $x = 0$. Expanding around one minimum $x_0 = +v$ by writing $\hat{x} = v + \hat{\eta}$:

$$V(v + \hat{\eta}) = V(v) + \frac{1}{2}(2\mu^2)\hat{\eta}^2 + O(\hat{\eta}^3). \quad (6.72)$$

The fluctuation field $\hat{\eta}$ sees a positive mass-squared $m^2 = 2\mu^2$: a stable oscillator. The “tachyonic” mass term $-\mu^2 \hat{x}^2/2$ disappears once you expand around the correct vacuum.

The ground state of the full H is a Fock state with a condensate: $\langle \Omega | \hat{x} | \Omega \rangle = v \neq 0$. The $\mathbb{Z}_2$ symmetry $\hat{x} \rightarrow -\hat{x}$ is spontaneously broken—not by the Hamiltonian (which is symmetric) but by the ground state (which picks one minimum). In QFT, this is the Higgs mechanism: v is the vacuum expectation value (VEV), $\hat{\eta}$ is the physical Higgs boson, and the mass $m = \sqrt{2}\mu$ is the Higgs mass generated by the curvature of the potential at the minimum.

6.10 Squeezed states

A squeezed state achieves the minimum uncertainty product $\Delta x \Delta p = 1/2$ but with the uncertainty distributed unequally: one quadrature is reduced below the coherent-state value at the cost of the other being enlarged.

The squeeze operator $S(\xi) = \exp\left(\frac{1}{2}\xi^* a^2 - \frac{1}{2}\xi a^{\dagger 2}\right)$ with $\xi = r e^{i\theta}$ acts on the ladder operators as a Bogoliubov transformation:

$$S^\dagger a S = a \cosh r - a^\dagger e^{i\theta} \sinh r. \quad (6.73)$$

The squeezed vacuum $|\xi\rangle = S(\xi)|0\rangle$ has variances:

$$(\Delta x)^2 = \frac{e^{-2r}}{2\omega}, \quad (\Delta p)^2 = \frac{\omega e^{+2r}}{2}. \quad (6.74)$$

Position is squeezed by e^{-r} ; momentum is anti-squeezed by e^{+r} . The product $\Delta x \Delta p = 1/2$ is preserved.

In the Fock basis, the squeezed vacuum contains only even excitations: $|\xi\rangle = (\cosh r)^{-1/2} \sum_{n=0}^{\infty} \frac{\sqrt{(2n)!}}{2^n n!} (-e^{i\theta} \tanh r)^n |2n\rangle$.

Applications. The Bogoliubov transformation (6.73) is the same transformation that connects the free vacuum to the BCS superconducting vacuum: Cooper pairing is squeezing. Squeezed light is used in LIGO to beat the quantum noise limit. Inflationary cosmology produces squeezed states: the inflaton mode amplitudes are squeezed by the exponential expansion, and the near-delta Born distribution of the squeezed quadrature is what “quantum fluctuations becoming classical perturbations” actually means.

6.11 The harmonic oscillator as the atom of quantum field theory

From one oscillator to infinitely many. The classical real scalar field $\phi(\mathbf{x}, t)$ satisfies the Klein–Gordon equation $(\partial_t^2 - \nabla^2 + m^2)\phi = 0$. Decomposing into spatial Fourier modes: $\phi(\mathbf{x}, t) = \int \frac{d^3 k}{(2\pi)^3} \tilde{\phi}(\mathbf{k}, t) e^{i\mathbf{k}\cdot\mathbf{x}}$, each mode satisfies $\ddot{\tilde{\phi}} + \omega_k^2 \tilde{\phi} = 0$ with $\omega_k = \sqrt{k^2 + m^2}$: a harmonic oscillator with frequency ω_k . The field is a collection of oscillators, one per mode, with frequency set by the relativistic dispersion relation.

Quantisation: promote amplitudes to operators. The classical amplitudes $c(\mathbf{k})$, $c^*(\mathbf{k})$ become operators $a(\mathbf{k})$, $a^\dagger(\mathbf{k})$ with $[a(\mathbf{k}), a^\dagger(\mathbf{k}')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}')$. The quantum field:

$$\hat{\phi}(\mathbf{x}, t) = \int \frac{d^3 k}{(2\pi)^3 \sqrt{2\omega_k}} [a(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x} - i\omega_k t} + a^\dagger(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x} + i\omega_k t}]. \quad (6.75)$$

The free field Hamiltonian.

$$H_{\text{free}} = \int \frac{d^3 k}{(2\pi)^3} \omega_k \left[a^\dagger(\mathbf{k}) a(\mathbf{k}) + \frac{1}{2} (2\pi)^3 \delta^3(0) \right]. \quad (6.76)$$

The second term is the zero-point energy: the sum of $\omega_k/2$ over all modes, which diverges in the continuum.

Normal ordering. The zero-point energy is a constant—it shifts all energy levels equally and produces no Born-layer consequences (all energy *differences* are unchanged). Normal ordering : H : subtracts this constant by placing all $a^\dagger$ ’s to the left of all a ’s:

$$: H_{\text{free}} := \int \frac{d^3 k}{(2\pi)^3} \omega_k a^\dagger(\mathbf{k}) a(\mathbf{k}). \quad (6.77)$$

This is the number operator weighted by ω_k : the energy is the sum of ω_k times the number of quanta in mode $\mathbf{k}$.

In the lattice-first viewpoint, the zero-point energy is an artefact of the continuum limit: on a finite lattice with spacing a , the sum over modes is finite, and the zero-point energy is a finite,

a -dependent constant that shifts the vacuum energy without affecting Born-layer observables. Normal ordering removes this a -dependent constant.

Adding interactions. The interacting scalar field Hamiltonian:

$$H =: H_{\text{free}} : + \frac{\lambda}{4!} \int d^3x : \hat{\phi}(\mathbf{x})^4 : \quad (6.78)$$

is a polynomial of degree 4 in $a(\mathbf{k})$ and $a^\dagger(\mathbf{k})$ —the same architectural form as the single-oscillator interacting motor (6.36), now with one pair of ladder operators per mode. The ϕ^4 vertex creates and destroys quanta across modes. The Standard Model Hamiltonian H_{SM} is a polynomial of the same kind, with creation and annihilation operators for every particle species (quarks, leptons, gauge bosons, Higgs), constrained by $SU(3)_c \times SU(2)_L \times U(1)_Y$ gauge invariance and Lorentz invariance.

The architecture is the same at every level: a polynomial motor of a and $a^\dagger$, applied to the Fock state at each tick. The harmonic oscillator is the atom from which the entire edifice is built.

7 The Feynman Path Integral

The path integral is a method for computing Fock-layer matrix elements of e^{-iHT} . It lives entirely inside the Fock layer. It does not replace the Schrödinger equation; it re-expresses one matrix element of e^{-iHT} as a high-dimensional integral that can be evaluated by Gaussian and perturbative techniques. The result is a number—a Fock-layer matrix element—not an observable. The bridge from these matrix elements to what the observer actually sees (Born-layer observables) requires additional machinery, which we develop at the end of this section.

7.1 The propagator as a Fock-layer matrix element

The central object is the *propagator*: the matrix element of e^{-iHT} between two position eigenstates:

$$K(x_f, T; x_i, 0) = \langle x_f | e^{-iHT} | x_i \rangle. \quad (7.1)$$

This is the (x_f, x_i) entry of the matrix e^{-iHT} in the position basis $\{|x\rangle\}$. On a discrete lattice with N_x grid points, e^{-iHT} is a literal $N_x \times N_x$ complex matrix, and $K(x_f, T; x_i, 0)$ is the entry in row x_f , column x_i . The propagator is a matrix element, and computing it is a matrix computation.

More generally, starting from any initial state $|\psi(0)\rangle$:

$$\psi(x_f, T) = \langle x_f | e^{-iHT} |\psi(0)\rangle = \int dx_i K(x_f, T; x_i, 0) \psi(x_i, 0), \quad (7.2)$$

where we inserted $\mathbf{1} = \int dx_i |x_i\rangle \langle x_i|$. The propagator is the kernel of the time evolution: it tells how Fock-layer amplitude at x_i flows to x_f .

The propagator is a Fock-layer quantity. Note what the propagator is *not*. It is not a Born-layer observable. No observer submits $K(x_f, T; x_i, 0)$ to the Born interface. It is not the probability that the system “goes from x_i to x_f ”—that would be the Born probability $|K(x_f, T; x_i, 0)|^2$, which requires a position measurement at time T (a Born-layer act). The propagator itself is a complex number inside the engine, awaiting further processing before it becomes a prediction.

7.2 Time slicing: matrix multiplication, not history

Split e^{-iHT} into N factors of $e^{-iH\varepsilon}$ with $\varepsilon = T/N$:

$$e^{-iHT} = \underbrace{e^{-iH\varepsilon} \cdots e^{-iH\varepsilon}}_{N \text{ factors}}. \quad (7.3)$$

Insert $\mathbf{1} = \int dx |x\rangle \langle x|$ between each pair:

$$K(x_f, T; x_i, 0) = \int dx_1 \cdots dx_{N-1} \prod_{k=0}^{N-1} \langle x_{k+1} | e^{-iH\varepsilon} | x_k \rangle = (U_\varepsilon^N)_{x_f, x_i}. \quad (7.4)$$

Each $K_\varepsilon(x_{k+1}, x_k) = \langle x_{k+1} | e^{-iH\varepsilon} | x_k \rangle$ is one entry of the short-time matrix U_ε in the position basis. The full propagator is the (x_f, x_i) entry of the N -th matrix power U_ε^N .

The integration over $x_1, \dots, x_{N-1}$ is matrix multiplication: summing over all intermediate basis indices is how you multiply matrices. The integral has the same mathematical structure as a sum over sequences $(x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_N)$, and it is common to call each such sequence a “path.” But the x_k at each time slice is a *bookkeeping eigenvalue*—a basis label inserted to decompose a matrix product. It is not a position the system “visits.” No observer measures x_k . No Born

rendering occurs at t_k . No Bohr trimming resets the state. The system does not take this path or any path. The hidden state is one vector, rotating under e^{-iHT} , and we are computing one entry of the rotation matrix by decomposing the computation into N steps.

Different choices of intermediate basis (energy, momentum) give different but equivalent decompositions of the same matrix product. The “sum over paths” is a coordinate-system choice for matrix multiplication, not a physical narrative.

7.3 The short-time matrix element and the Lagrangian

For small ε , the matrix element $K_\varepsilon(x_{k+1}, x_k) = \langle x_{k+1} | e^{-iH\varepsilon} | x_k \rangle$ can be evaluated explicitly. This is the single most important calculation in the path integral formalism: it is where the Lagrangian re-emerges from the Hamiltonian—not as a postulate, but as a consequence of a Gaussian integral.

Step 1: insert the momentum resolution of the identity. With $H = \hat{P}^2/2m + V(\hat{X})$ and $\mathbf{1} = \int dp/(2\pi) |p\rangle \langle p|$:

$$\begin{aligned} K_\varepsilon(x_{k+1}, x_k) &= \langle x_{k+1} | e^{-iH\varepsilon} | x_k \rangle \\ &\approx \langle x_{k+1} | e^{-i\hat{P}^2\varepsilon/2m} e^{-iV(\hat{X})\varepsilon} | x_k \rangle + O(\varepsilon^2), \end{aligned} \quad (7.5)$$

where we used the Trotter splitting $e^{-i(A+B)\varepsilon} \approx e^{-iA\varepsilon} e^{-iB\varepsilon} + O(\varepsilon^2)$. (The error is $O(\varepsilon^2)$ because $[A, B] \neq 0$, but it vanishes in the $N \rightarrow \infty$ limit.)

Since $|x_k\rangle$ is an eigenstate of $\hat{X}$: $e^{-iV(\hat{X})\varepsilon} |x_k\rangle = e^{-iV(x_k)\varepsilon} |x_k\rangle$. Insert $\mathbf{1} = \int dp/(2\pi) |p\rangle \langle p|$:

$$\begin{aligned} K_\varepsilon(x_{k+1}, x_k) &= \int \frac{dp}{2\pi} \langle x_{k+1} | p \rangle \langle p | e^{-i\hat{P}^2\varepsilon/2m} | x_k \rangle e^{-iV(x_k)\varepsilon} \\ &= \int \frac{dp}{2\pi} \frac{e^{ipx_{k+1}}}{\sqrt{2\pi}} \cdot e^{-ip^2\varepsilon/2m} \cdot \frac{e^{-ipx_k}}{\sqrt{2\pi}} \cdot e^{-iV(x_k)\varepsilon} \\ &= \int \frac{dp}{2\pi} e^{ip(x_{k+1}-x_k)} e^{-ip^2\varepsilon/(2m)} e^{-iV(x_k)\varepsilon}. \end{aligned} \quad (7.6)$$

We used $\langle x|p\rangle = e^{ipx}/\sqrt{2\pi}$ and $\hat{P}|p\rangle = p|p\rangle$.

Step 2: Gaussian integration over p (completing the square). The p -integral in (7.6) is:

$$I = \int \frac{dp}{2\pi} \exp\left(-\frac{i\varepsilon}{2m}p^2 + ip\Delta x\right), \quad \Delta x \equiv x_{k+1} - x_k. \quad (7.7)$$

This is a Gaussian integral in p . Complete the square in the exponent:

$$\begin{aligned} -\frac{i\varepsilon}{2m}p^2 + ip\Delta x &= -\frac{i\varepsilon}{2m}\left(p^2 - \frac{2m\Delta x}{\varepsilon}p\right) \\ &= -\frac{i\varepsilon}{2m}\left(p - \frac{m\Delta x}{\varepsilon}\right)^2 + \frac{im(\Delta x)^2}{2\varepsilon}. \end{aligned} \quad (7.8)$$

The cross term cancels perfectly. Substituting back:

$$I = e^{im(\Delta x)^2/(2\varepsilon)} \int \frac{dp}{2\pi} \exp\left(-\frac{i\varepsilon}{2m}\left(p - \frac{m\Delta x}{\varepsilon}\right)^2\right). \quad (7.9)$$

Shift the integration variable $p \rightarrow p + m\Delta x/\varepsilon$ (this does not change the limits because p runs from $-\infty$ to $+\infty$):

$$I = e^{im(\Delta x)^2/(2\varepsilon)} \underbrace{\int \frac{dp}{2\pi} e^{-i\varepsilon p^2/(2m)}}_{=\sqrt{m/(2\pi i\varepsilon)}}. \quad (7.10)$$

The remaining Gaussian integral is $\int_{-\infty}^{\infty} dp e^{-\alpha p^2} = \sqrt{\pi/\alpha}$ with $\alpha = i\varepsilon/(2m)$, giving $\sqrt{\pi \cdot 2m/(i\varepsilon)} = \sqrt{2\pi m/(i\varepsilon)}$. Dividing by 2π : $\sqrt{m/(2\pi i\varepsilon)}$.

Step 3: the result—the Lagrangian appears. Combining:

$$K_\varepsilon(x_{k+1}, x_k) = \sqrt{\frac{m}{2\pi i\varepsilon}} \exp\left(i\varepsilon \left[\frac{m}{2} \left(\frac{\Delta x}{\varepsilon} \right)^2 - V(x_k) \right]\right). \quad (7.11)$$

The exponent is $i\varepsilon \mathcal{L}(x_k, \dot{x}_k)$, where:

$$\mathcal{L}(x, \dot{x}) = \frac{m\dot{x}^2}{2} - V(x), \quad \dot{x}_k \approx \frac{x_{k+1} - x_k}{\varepsilon}, \quad (7.12)$$

is the classical Lagrangian: kinetic energy *minus* potential energy.

What happened. The Hamiltonian $H = p^2/(2m) + V(x)$ enters the matrix element. The Gaussian integration over the momentum p trades the quadratic $p^2/(2m)$ for the quadratic $m\dot{x}^2/2$ —the Legendre transform from (x, p) to $(x, \dot{x})$, executed by completing the square. The result is the Lagrangian $\mathcal{L} = m\dot{x}^2/2 - V(x)$ in the exponent.

The Lagrangian is not a postulate. It is the result of a Gaussian integral inside the Fock layer. The short-time matrix element of $e^{-iH\varepsilon}$ in the position basis happens to have the form $e^{i\varepsilon\mathcal{L}}$, because integrating out the momentum from the Hamiltonian produces the Lagrangian by the same Legendre-transform step that connects them classically.

This is why the Lagrangian appears in quantum mechanics at all: it is the Hamiltonian with the momentum integrated out—a Fock-layer computational artefact, not a foundational principle.

7.4 The path integral formula

Substituting (7.11) into (7.4) and taking $N \rightarrow \infty$:

$$K(x_f, T; x_i, 0) = \int_{x(0)=x_i}^{x(T)=x_f} \mathcal{D}x(t) e^{iS[x]}, \quad S[x] = \int_0^T \mathcal{L}(x, \dot{x}) dt. \quad (7.13)$$

The notation $\int \mathcal{D}x$ means the limit of the $(N-1)$ -fold integral $\int dx_1 \cdots dx_{N-1}$ as $N \rightarrow \infty$. $S[x]$ is the classical action of the configuration $x(t)$. Each configuration contributes a complex amplitude $e^{iS[x]}$ of unit modulus.

The Fock-Born diagnostic. The field value $x(t_k)$ at each time slice is a bookkeeping eigenvalue: a basis label inserted to decompose a Fock-layer matrix element. It is *not* the outcome of a Born-layer measurement. No observer samples it. No Bohr trimming occurs at t_k . The system does not “visit” the configuration $x(t_k)$.

The standard name “sum over histories” invites the reader to interpret each configuration $x(t)$ as a history the system takes—a sequence of Born-layer actualities. This is the conflation catalogued in the preface: Born-layer language pushed into the Fock layer. The “histories” are integration variables in a matrix product. The system takes no history. The hidden state is one vector, rotating under e^{-iHT} , and the path integral computes one entry of the rotation matrix.

For fermions, the integration variables are Grassmann numbers—formal anticommuting symbols with no classical meaning. There is no fermionic field configuration the system “visits.” This should serve as a standing reminder that path-integral variables are computational devices, not physical states—even in the bosonic case, where the temptation to reify them is stronger because the eigenvalues happen to be real numbers.

7.5 Stationary phase and the classical trajectory

Each configuration contributes $e^{iS[x]}$, a complex number of unit modulus. For most configurations, neighbouring ones have actions that differ by much more than 1 (in units with $\hbar = 1$), so their phases point in different directions and cancel when summed. The only configurations that survive the cancellation are those near the stationary point of the action: $\delta S = 0$ for all small variations $\delta x(t)$.

The stationary-phase condition gives the Euler–Lagrange equation:

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}} - \frac{\partial \mathcal{L}}{\partial x} = 0 \quad \implies \quad m\ddot{x} = -V'(x) : \quad (7.14)$$

Newton’s second law. The classical trajectory $x_{\text{cl}}(t)$ is the configuration of stationary phase. Near the classical trajectory, the action varies slowly; away from it, the phases oscillate wildly and cancel. When $S \gg \hbar$, only the classical trajectory and its immediate neighbourhood contribute appreciably to the matrix element.

What stationary phase does and does not explain. Stationary phase tells you which Fock-layer matrix elements $\langle x_f | e^{-iHT} | x_i \rangle$ are large: those where the boundary conditions (x_i, x_f, T) are connected by a classical trajectory. This is a powerful computational result—it reduces an infinite-dimensional integral to the neighbourhood of a single configuration.

But stationary phase does *not* explain the emergence of the classical world. In the Fock-Born framework, the classical world emerges through the Born layer: when a macroscopic observable has a near-delta Born distribution (variance $\sim 1/N$, degeneracy $\sim e^S$), Born sampling is deterministic and Bohr trimming is inert. The observer sees a definite value at every measurement because the distribution has already decided, not because off-classical configurations cancel in a path integral.

Stationary phase is a useful approximation technique inside the Fock layer. The emergence of classicality is a property of the Born layer. These are different statements in different layers of the architecture, and conflating them—treating stationary phase as the explanation of classicality—is a layer violation.

7.6 The harmonic oscillator path integral

For $V(x) = \omega^2 x^2/2$ (with $m = 1$), the Lagrangian is $\mathcal{L} = \dot{x}^2/2 - \omega^2 x^2/2$ and the action is quadratic in $x(t)$. This means the path integral is Gaussian and can be evaluated exactly.

Step 1: the classical trajectory. The classical equation of motion is $\ddot{x}_{\text{cl}} + \omega^2 x_{\text{cl}} = 0$, with boundary conditions $x_{\text{cl}}(0) = x_i$ and $x_{\text{cl}}(T) = x_f$. The general solution is $x_{\text{cl}}(t) = A \cos \omega t + B \sin \omega t$. Imposing the boundary conditions:

$$x_{\text{cl}}(0) = A = x_i, \quad (7.15)$$

$$x_{\text{cl}}(T) = x_i \cos \omega T + B \sin \omega T = x_f \quad \Rightarrow \quad B = \frac{x_f - x_i \cos \omega T}{\sin \omega T}. \quad (7.16)$$

Therefore:

$$x_{\text{cl}}(t) = \frac{x_i \sin \omega(T-t) + x_f \sin \omega t}{\sin \omega T}. \quad (7.17)$$

Check: $x_{\text{cl}}(0) = x_i \sin \omega T / \sin \omega T = x_i$; $x_{\text{cl}}(T) = x_f \sin \omega T / \sin \omega T = x_f$. ✓

Step 2: the classical action. The action is $S = \int_0^T (\dot{x}^2/2 - \omega^2 x^2/2) dt$. For a solution of the

equation of motion, a useful shortcut: integrate by parts.

$$\begin{aligned} S[x_{\text{cl}}] &= \int_0^T \left(\frac{\dot{x}_{\text{cl}}^2}{2} - \frac{\omega^2 x_{\text{cl}}^2}{2} \right) dt \\ &= \left[\frac{x_{\text{cl}} \dot{x}_{\text{cl}}}{2} \right]_0^T - \frac{1}{2} \int_0^T x_{\text{cl}} \underbrace{(\ddot{x}_{\text{cl}} + \omega^2 x_{\text{cl}})}_{=0} dt = \frac{1}{2} [x_{\text{cl}} \dot{x}_{\text{cl}}]_0^T. \end{aligned} \quad (7.18)$$

(The integral vanishes because x_{cl} satisfies the equation of motion.) Compute $\dot{x}_{\text{cl}}$:

$$\dot{x}_{\text{cl}}(t) = \frac{\omega[-x_i \cos \omega(T-t) + x_f \cos \omega t]}{\sin \omega T}. \quad (7.19)$$

Evaluate at the boundaries:

$$x_{\text{cl}}(T) \dot{x}_{\text{cl}}(T) = x_f \cdot \frac{\omega(-x_i + x_f \cos \omega T)}{\sin \omega T} = \frac{\omega(x_f^2 \cos \omega T - x_i x_f)}{\sin \omega T}, \quad (7.20)$$

$$x_{\text{cl}}(0) \dot{x}_{\text{cl}}(0) = x_i \cdot \frac{\omega(-x_i \cos \omega T + x_f)}{\sin \omega T} = \frac{\omega(x_i x_f - x_i^2 \cos \omega T)}{\sin \omega T}. \quad (7.21)$$

Subtract:

$$\boxed{S[x_{\text{cl}}] = \frac{\omega}{2 \sin \omega T} [(x_i^2 + x_f^2) \cos \omega T - 2x_i x_f]}. \quad (7.22)$$

Step 3: splitting the path integral. Write any configuration as $x(t) = x_{\text{cl}}(t) + \xi(t)$ with $\xi(0) = \xi(T) = 0$ (the fluctuation vanishes at the endpoints). The action splits:

$$S[x_{\text{cl}} + \xi] = S[x_{\text{cl}}] + S_{\text{cross}} + S[\xi]. \quad (7.23)$$

The cross term $S_{\text{cross}} = \int_0^T (\dot{x}_{\text{cl}} \dot{\xi} - \omega^2 x_{\text{cl}} \xi) dt$ vanishes: integrate by parts to get $-\int_0^T (\ddot{x}_{\text{cl}} + \omega^2 x_{\text{cl}}) \xi dt = 0$ (by the equation of motion), plus boundary terms $[x_{\text{cl}} \dot{\xi}]_0^T$ which vanish because $\xi(0) = \xi(T) = 0$ does not constrain $\dot{\xi}$ at the boundaries—rather, the integration-by-parts boundary term is $[\dot{x}_{\text{cl}} \xi]_0^T = 0$ since $\xi(0) = \xi(T) = 0$.

The action splits exactly:

$$S[x_{\text{cl}} + \xi] = S[x_{\text{cl}}] + S[\xi]. \quad (7.24)$$

The classical action is pulled out of the path integral:

$$K_{\text{HO}}(x_f, T; x_i, 0) = e^{iS[x_{\text{cl}}]} \underbrace{\int_{\xi(0)=0}^{\xi(T)=0} \mathcal{D}\xi e^{iS[\xi]}}_{F(T)}. \quad (7.25)$$

Step 4: the fluctuation determinant. The fluctuation integral $F(T)$ is Gaussian: $S[\xi] = \frac{1}{2} \int_0^T (\dot{\xi}^2 - \omega^2 \xi^2) dt$ is quadratic in ξ . To evaluate it, expand $\xi(t)$ in the eigenmodes of the differential operator $-d^2/dt^2 - \omega^2$ on $[0, T]$ with Dirichlet boundary conditions:

$$\xi(t) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi t}{T}. \quad (7.26)$$

These modes are orthogonal: $\int_0^T \sin(n\pi t/T) \sin(m\pi t/T) dt = (T/2) \delta_{nm}$. The action in terms of the mode coefficients:

$$S[\xi] = \frac{T}{4} \sum_{n=1}^{\infty} \left(\frac{n^2 \pi^2}{T^2} - \omega^2 \right) b_n^2 = \frac{T}{4} \sum_{n=1}^{\infty} \lambda_n b_n^2, \quad (7.27)$$

where $\lambda_n = n^2\pi^2/T^2 - \omega^2$ is the n -th eigenvalue. Each mode contributes an independent Gaussian integral:

$$F(T) = \prod_{n=1}^{\infty} \int db_n \exp\left(\frac{iT\lambda_n}{4} b_n^2\right) \propto \prod_{n=1}^{\infty} \frac{1}{\sqrt{\lambda_n}}. \quad (7.28)$$

The infinite product of eigenvalues is evaluated using the product formula for the sine function:

$$\prod_{n=1}^{\infty} \left(1 - \frac{\omega^2 T^2}{n^2 \pi^2}\right) = \frac{\sin \omega T}{\omega T}, \quad (7.29)$$

which gives (including the prefactor from the free-particle case $\omega = 0$):

$$F(T) = \sqrt{\frac{\omega}{2\pi i \sin \omega T}}. \quad (7.30)$$

Step 5: the exact propagator. Combining:

$$K_{\text{HO}}(x_f, T; x_i, 0) = \sqrt{\frac{\omega}{2\pi i \sin \omega T}} \exp\left(\frac{i\omega}{2 \sin \omega T} [(x_i^2 + x_f^2) \cos \omega T - 2x_i x_f]\right). \quad (7.31)$$

For $\omega \rightarrow 0$: $K_{\text{free}} = (2\pi i T)^{-1/2} e^{i(x_f - x_i)^2/(2T)}$. The prefactor diverges at $\omega T = n\pi$ (focal points/caustics, where the classical trajectory is not unique).

7.7 Perturbation theory, Wick's theorem, and Feynman diagrams

Add a small nonlinear perturbation: $V(x) = \omega^2 x^2/2 + \lambda x^4/4!$, $|\lambda| \ll 1$. The path integral becomes:

$$K = \int \mathcal{D}x \, e^{iS_{\text{HO}}[x]} e^{-i\frac{\lambda}{4!} \int x(t)^4 dt}. \quad (7.32)$$

Taylor-expand the interaction exponential in powers of λ :

$$K = K_{\text{HO}} - \frac{i\lambda}{4!} \int dt_1 \langle x(t_1)^4 \rangle_{\text{HO}} \cdot K_{\text{HO}} + O(\lambda^2), \quad (7.33)$$

where $\langle \cdots \rangle_{\text{HO}}$ denotes the average over the Gaussian (harmonic oscillator) path-integral measure.

Wick's theorem: derivation and enumeration

Statement. For any Gaussian measure with covariance (“propagator”) $G(t, t') = \langle x(t)x(t') \rangle_{\text{HO}}$, the expectation of a product of $2n$ fields factors into a sum over all complete pairings:

$$\langle x(t_1)x(t_2)\cdots x(t_{2n}) \rangle_{\text{HO}} = \sum_{\text{pairings}} \prod_{\text{pairs } (i,j)} G(t_i, t_j). \quad (7.34)$$

The sum runs over all ways of grouping the $2n$ labels into n pairs. The number of pairings is $(2n-1)!! = (2n-1)(2n-3)\cdots 3\cdot 1$. Odd moments vanish: $\langle x(t_1)\cdots x(t_{2n+1}) \rangle = 0$.

Proof for $n = 2$ (four fields). The key tool is the generating function: for a Gaussian measure with covariance G , the moment-generating function is:

$$\langle e^{i \sum_a J_a x(t_a)} \rangle_{\text{HO}} = \exp\left(-\frac{1}{2} \sum_{a,b} J_a G(t_a, t_b) J_b\right). \quad (7.35)$$

(This is the standard Gaussian identity $\langle e^{iJx} \rangle = e^{-J^T G J/2}$ for a multivariate Gaussian.)

To extract $\langle x_1 x_2 x_3 x_4 \rangle$ (writing $x_a \equiv x(t_a)$), differentiate four times:

$$\langle x_1 x_2 x_3 x_4 \rangle = \frac{1}{i^4} \frac{\partial^4}{\partial J_1 \partial J_2 \partial J_3 \partial J_4} e^{-\frac{1}{2} \sum_{a,b} J_a G_{ab} J_b} \Big|_{J=0}. \quad (7.36)$$

Differentiate $e^{-J^T G J/2}$ with respect to J_1 : this brings down a factor $-\sum_b G_{1b} J_b$. Differentiate again with respect to J_2 : the derivative hits either the exponential (bringing down $-\sum_c G_{2c} J_c$) or the pre-existing factor (giving $-G_{12}$). Continuing through all four derivatives and setting $J = 0$ (which kills all terms with residual J factors), the surviving terms are exactly the three complete pairings:

$$\boxed{\langle x_1 x_2 x_3 x_4 \rangle = G_{12} G_{34} + G_{13} G_{24} + G_{14} G_{23}.} \quad (7.37)$$

The three pairings are:

Pairing	Pairs	Contribution
(12)(34)	1–2 and 3–4	$G(t_1, t_2) G(t_3, t_4)$
(13)(24)	1–3 and 2–4	$G(t_1, t_3) G(t_2, t_4)$
(14)(23)	1–4 and 2–3	$G(t_1, t_4) G(t_2, t_3)$

For the special case $t_1 = t_2 = t_3 = t_4$ (all at the same point):

$$\langle x(t)^4 \rangle = 3 G(t, t)^2. \quad (7.38)$$

The factor $3 = (4 - 1)!! = 3!!$ counts the pairings of 4 objects.

Wick’s theorem for general n . The same differentiation argument extends to $2n$ fields: each differentiation $\partial/\partial J_a$ either brings down a new factor $-\sum_b G_{ab} J_b$ from the exponential or contracts with an existing factor by replacing J_b with G_{ab} . After $2n$ differentiations and setting $J = 0$, the only surviving terms are those where every J has been paired—exactly the complete pairings of $2n$ labels into n pairs. The number of such pairings is $(2n)!/(2^n n!) = (2n - 1)!!$.

In ML terms: Wick’s theorem says that the statistics of a Gaussian distribution are completely determined by its covariance matrix. This is the statement that a Gaussian has no irreducible higher-order correlations—all moments factor into products of second moments.

Feynman diagrams as terms in the expansion

Each term in (7.33) has a diagrammatic representation: lines for propagators $G(t, t')$, vertices for factors of $-i\lambda/4!$, integrated over vertex times.

At order λ^1 : the $\langle x(t_1)^4 \rangle$ factor produces $3 G(t_1, t_1)^2$ —three pairings, each a product of two propagators that start and end at the same point t_1 . Diagrammatically: a vertex at t_1 with two closed loops (“figure eight” or “vacuum bubble”).

At order λ^2 : $\langle x(t_1)^4 x(t_2)^4 \rangle$ produces $9 \cdot 7 \cdot 5 \cdot 3 \cdot 1 = 105$ pairings of 8 fields, but many are disconnected (factorise into products of lower-order diagrams). The connected diagrams at order λ^2 include the “sunset” (two propagators connecting t_1 to t_2 with a third looping back) and the “double bubble” (two self-contractions, one at each vertex).

Every element of the diagram is a Fock-layer bookkeeping device:

A *line* is a Green’s function $G(t, t')$ —a matrix element of the inverse of the quadratic operator $-\partial_t^2 - \omega^2$. It is the free propagator: the amplitude for the engine to evolve a fluctuation from time t' to time t . It is not a particle travelling from t' to t .

A *vertex* is a factor $-i\lambda/4!$ from the Taylor expansion of $e^{-iS_{\text{int}}}$ —one application of the interaction polynomial. It is not an event where particles are created or destroyed. It is a term in a power series.

A *closed loop* is a self-contraction $G(t_1, t_1)$ —a propagator from a point to itself. It is not a “virtual particle” popping into and out of existence. It is a matrix element evaluated at coincident arguments.

The perturbative expansion of the path integral and the Dyson series of section 6.7 produce exactly the same diagrams, the same amplitudes, and the same symmetry factors. The Dyson series works in operator language ($a^\dagger$, a , Fock-state matrix elements); the path integral works in eigenvalue language ($x(t)$, Gaussian integrals, Green’s functions). Both are Fock-layer computations that decompose e^{-iHT} into manageable pieces. The Feynman diagram is the shared notation for both decompositions.

The sum-over-histories narrative: a systematic refutation

The path integral is commonly described as a “sum over all histories the system might take,” with each history $x(t)$ weighted by $e^{iS[x]}$. This language is vivid and pedagogically useful, but it conflates the two layers of the Fock-Born architecture. We now refute the narrative point by point.

Claim: “The system takes every possible path from x_i to x_f .”

Refutation: The system does not take any path. The hidden state $h_t = e^{-iHt}h_0$ is one vector in Hilbert space, rotating under the Hamiltonian. The path integral computes one matrix element of this rotation— $\langle x_f | e^{-iHT} | x_i \rangle$ —by decomposing the matrix product into intermediate basis insertions. The “paths” $x(t_0), x(t_1), \dots, x(t_N)$ are the intermediate basis labels—the row and column indices summed over in a matrix multiplication. They are not positions the system visits. No observer measures $x(t_k)$ at the intermediate time t_k . No Born rendering occurs. No Bohr trimming resets the state.

Claim: “Each path has a probability amplitude $e^{iS[x]}$.”

Refutation: Each “path” contributes a factor $\prod_k K_\varepsilon(x_{k+1}, x_k)$ to the matrix element. This is a product of matrix entries along a sequence of intermediate indices—the same algebraic structure as summing over intermediate states in a matrix power $A_{ij}^N = \sum_{k_1, \dots} A_{ik_1} A_{k_1 k_2} \cdots A_{k_{N-1} j}$. The word “amplitude” invites the interpretation that it is a Born-layer probability amplitude—a quantity whose squared modulus is a detection probability. It is not. It is a term in a matrix computation. The Born-layer probability is $|K(x_f, T; x_i, 0)|^2$, which requires a Born-layer measurement at time T and is a property of the *sum* over all paths, not of any individual path.

Claim: “The classical path is the one with the largest contribution.”

Refutation: The classical path is the stationary- phase configuration $\delta S = 0$. Near this path, the phases e^{iS} point in approximately the same direction and add constructively. Far from this path, the phases oscillate wildly and cancel. This is a useful computational observation about which matrix elements are large—it is a property of the Fock-layer computation, not a statement about what the system does. The emergence of the classical world is a Born-layer phenomenon (the Born distribution concentrates via degeneracy and the law of large numbers), not a path-integral phenomenon (constructive interference at the stationary phase).

The Grassmann reminder. For fermions, the path-integral variables are Grassmann numbers—formal anticommuting symbols with no classical interpretation. There is no fermionic field configuration the system “visits.” The Grassmann integral $\int d\bar{\psi} d\psi e^{-\bar{\psi} M \psi} = \det M$ is an alge-

braic identity for computing a determinant, not a sum over fermionic histories. This should serve as a permanent reminder that path-integral variables are computational devices, not physical states—even in the bosonic case, where the temptation to reify them is stronger because the variables happen to be real numbers.

The correct language.

Common language (conflation)	Fock-Born language (precise)
“The system takes every path”	The matrix element is computed by summing over intermediate basis labels
“Each path has an amplitude e^{iS} ”	Each sequence of basis labels contributes a product of matrix entries
“The paths interfere”	The matrix entries are complex and can cancel when summed
“The classical path dominates”	The stationary-phase matrix element is largest; off-stationary contributions cancel
“Sum over histories”	Matrix power formula: $\langle x_f U^N x_i \rangle = \sum_{k_1 \dots} U_{x_f k_1} U_{k_1 k_2} \dots U_{k_{N-1} x_i}$

The path integral is a powerful computational tool. It is also a powerful source of conflation, because its language (“paths,” “histories,” “amplitudes”) belongs to the Born layer while its content belongs to the Fock layer. The cure is not to avoid the path integral—it is far too useful for that—but to remember, at every step, that the “path” is a basis label, the “amplitude” is a matrix element, and the “history” is an index in a sum.

7.8 The generating function and the bridge to Born-layer observables

The path integral computes Fock-layer matrix elements. The observer measures Born-layer observables. This subsection builds the bridge.

The generating function $Z[J]$. Introduce an external source $J(t)$ —a classical function coupled linearly to the field:

$$Z[J] = \int \mathcal{D}x \exp\left(i \int_0^T [\mathcal{L}(x, \dot{x}) + J(t) x(t)] dt\right) = \int \mathcal{D}x e^{i(S[x] + \int J x dt)}. \quad (7.39)$$

In the operator language, $Z[J]$ is the vacuum-to-vacuum amplitude in the presence of the source:

$$Z[J] = \langle \Omega | T \exp\left(i \int dt J(t) \hat{x}(t)\right) | \Omega \rangle, \quad (7.40)$$

where $\hat{x}(t)$ is the Heisenberg-picture position operator. The source $J(t)$ is a handle that lets us probe the engine from outside: by differentiating $Z[J]$ with respect to J , we extract correlation functions.

Correlation functions by differentiation. Functional derivatives of $Z[J]$ with respect to J generate time-ordered correlation functions:

$$\left. \frac{1}{i} \frac{\delta}{\delta J(t_1)} Z[J] \right|_{J=0} = \langle \Omega | \hat{x}(t_1) | \Omega \rangle, \quad (7.41)$$

$$\left. \frac{1}{i^2} \frac{\delta^2}{\delta J(t_1) \delta J(t_2)} Z[J] \right|_{J=0} = \langle \Omega | T \{ \hat{x}(t_1) \hat{x}(t_2) \} | \Omega \rangle. \quad (7.42)$$

The n -point function $\langle \Omega | T \{ \hat{x}(t_1) \cdots \hat{x}(t_n) \} | \Omega \rangle$ is obtained by n functional derivatives. The connected correlation functions come from $W[J] = \ln Z[J]$, and the one-particle-irreducible (1PI) vertex functions come from the Legendre transform $\Gamma[\bar{x}]$.

All of this is inside the Fock layer. $Z[J]$, $W[J]$, $\Gamma[\bar{x}]$, and the correlation functions are Fock-layer quantities: matrix elements of time-ordered products of operators, computed by Gaussian integrals and perturbative expansions. They encode the complete dynamical information of the engine—the spectrum, the couplings, the vacuum structure—but they are not themselves Born-layer observables. No observer submits a “two-point function” to the Born interface.

The generating function is the engine’s logbook: a complete record of how the engine responds to external probes. The correlation functions are the logbook’s entries. The question now is: how does the observer read this logbook?

Bridge 1: Born-layer expectation values. The simplest connection. If the observer measures the position observable $\hat{x}$ at time t on the state $|\Omega\rangle$, the Born expected value is:

$$\langle \hat{x}(t) \rangle = \langle \Omega | \hat{x}(t) | \Omega \rangle = \left. \frac{1}{i} \frac{\delta}{\delta J(t)} \ln Z[J] \right|_{J=0}. \quad (7.43)$$

The Born variance is:

$$(\Delta x)^2(t) = \langle \Omega | \hat{x}(t)^2 | \Omega \rangle - \langle \Omega | \hat{x}(t) | \Omega \rangle^2 = \left. \frac{1}{i^2} \frac{\delta^2}{\delta J(t)^2} \ln Z[J] \right|_{J=0}. \quad (7.44)$$

The expected value and variance are Born-layer predictions: they describe the distribution the observer would see if she measured $\hat{x}$ on many identically prepared copies of $|\Omega\rangle$. They are extracted from $Z[J]$ —the Fock-layer generating function—by differentiation. The differentiation is the bridge.

Bridge 2: the spectrum (masses and energy levels). The eigenvalues of the full Hamiltonian H are Born-layer observables: they are what a spectrometer reports. The path integral encodes them through the large-time behaviour of the propagator.

In Euclidean time ($t \rightarrow -i\tau$):

$$K_E(x_f, \tau; x_i, 0) = \langle x_f | e^{-H\tau} | x_i \rangle = \sum_n \psi_n(x_f) \psi_n^*(x_i) e^{-E_n \tau}. \quad (7.45)$$

At large τ , the sum is dominated by the ground state ($n = 0$), and the energy gap $E_1 - E_0$ is extracted from the exponential decay rate:

$$\frac{K_E(\tau)}{K_E(\tau + \Delta\tau)} \xrightarrow{\tau \rightarrow \infty} e^{E_0 \Delta\tau}. \quad (7.46)$$

In QFT, the pole of the two-point function in momentum space gives the physical mass: $G(p) \sim 1/(p^2 - m_{\text{phys}}^2 + i\epsilon)$. The mass m_{phys} is a Born-layer quantity (measurable by a spectrometer); the propagator $G(p)$ is a Fock-layer intermediary from which the mass is extracted.

Bridge 3: scattering cross sections via the S -matrix and LSZ. The S -matrix element $\langle f | S | i \rangle = \langle f | T \exp(-i \int H_{\text{int}} dt) | i \rangle$ gives the amplitude for the transition from initial state $|i\rangle$ to final state $|f\rangle$. The Born probability is:

$$P(i \rightarrow f) = |\langle f | S | i \rangle|^2. \quad (7.47)$$

This is a Born-layer observable: the fraction of scattering events in which detectors register final state f , given that the accelerator prepared initial state i . It is what experimentalists publish as a cross section σ .

The **LSZ reduction formula** is the bridge between the Fock-layer correlation functions (computed by $Z[J]$) and the Born-layer S -matrix elements. Schematically:

$$\langle f | S | i \rangle = \prod_{\text{external}} (p_j^2 - m^2) \tilde{G}^{(n)}(p_1, \dots, p_n), \quad (7.48)$$

where $\tilde{G}^{(n)}$ is the Fourier-transformed n -point correlation function and the product of $(p_j^2 - m^2)$ factors “amputates” the external propagators. The LSZ formula converts Fock-layer correlation functions into Born-layer scattering amplitudes. It is the exit gate from the engine to the screen.

Bridge 4: decay widths, lifetimes, and spectral functions. The decay width Γ of an unstable state is extracted from the imaginary part of the self-energy $\Sigma(p)$:

$$G(p) = \frac{1}{p^2 - m_0^2 - \Sigma(p)} \approx \frac{1}{p^2 - m_{\text{phys}}^2 + im_{\text{phys}}\Gamma}, \quad (7.49)$$

where $m_{\text{phys}}^2 = m_0^2 + \text{Re } \Sigma(m_{\text{phys}}^2)$ and $\Gamma = -\text{Im } \Sigma(m_{\text{phys}}^2)/m_{\text{phys}}$. The lifetime $\tau = 1/\Gamma$ is a Born-layer observable (the time constant of the Weisskopf–Wigner exponential decay law); the self-energy $\Sigma(p)$ is a Fock-layer intermediary.

The spectral function $\rho(\omega) = -\frac{1}{\pi} \text{Im } G^R(\omega)$ gives the distribution of excitation energies in the interacting system. It is directly measurable by spectroscopy (Born-layer), and it encodes the full eigenvalue spectrum of H .

Bridge 5: thermodynamic quantities. At finite temperature, the partition function is:

$$\mathcal{Z}(\beta) = \text{Tr } e^{-\beta H} = \int dx K_E(x, \beta; x, 0) = \int \mathcal{D}x e^{-S_E[x]}, \quad (7.50)$$

where the Euclidean path integral is over periodic configurations $x(0) = x(\beta)$. Thermodynamic observables—free energy $F = -\beta^{-1} \ln \mathcal{Z}$, energy $\langle H \rangle = -\partial \ln \mathcal{Z} / \partial \beta$, specific heat C_V , susceptibilities—are all Born-layer quantities extracted from the Fock-layer object $\mathcal{Z}(\beta)$ by differentiation.

Summary: the two-column structure.

Fock-layer intermediaries (computed by path integral)	Born-layer observables (what the observer measures)
Propagator $K(x_f, T; x_i, 0)$	Transition probability $ K ^2$ (after Born rendering)
Generating function $Z[J]$	Expectation values $\langle \hat{\mathcal{O}} \rangle$ (by differentiation)
Correlation functions $\langle T \{ \hat{\phi}(x_1) \cdots \hat{\phi}(x_n) \} \rangle$	Scattering cross sections σ (via LSZ)
Self-energy $\Sigma(p)$	Mass m_{phys} , decay width Γ , lifetime τ
Spectral function $\rho(\omega)$	Spectral lines (spectrometer readings)
Running coupling $g(\mu)$	Cross section at energy μ
Euclidean partition function $\mathcal{Z}(\beta)$	Free energy F , specific heat C_V , susceptibility χ
Vertex functions, wave-function renormalisation	Cancel in physical predictions; never observed

The path integral computes the left column. The Born layer lives in the right column. The bridges (differentiation, LSZ, spectral analysis, thermodynamic derivatives) convert Fock-layer intermediaries into Born-layer predictions. The entire Feynman-diagram machinery—propagators, vertices, loops, Wick contractions, renormalisation—operates in the left column. No element of a Feynman diagram is a Born-layer event.

7.9 Path integral and Markov chain: amplitude versus probability

The path integral and the classical Markov chain share a mathematical skeleton: both compute the (q_f, q_i) entry of an N -th matrix power by summing over all intermediate-state sequences. The difference is a single word: Markov chains sum *probabilities* (positive reals); path integrals sum *amplitudes* (complex numbers of unit modulus).

	Markov chain	Path integral
One-step weight	$T_{q'q} \geq 0$ (probability)	$U_{q'q} \in \mathbb{C}$, $ U_{q'q} = 1$ (amplitude)
N -step result	$(T^N)_{q_f q_i}$ (probability)	$(U^N)_{q_f q_i}$ (amplitude)
Weight property	Positive, sums to 1	Complex, can cancel
Interference	No: probabilities add	Yes: amplitudes add, can cancel
Physical quantity	Probability P directly	$ \text{amplitude} ^2 = P$ (Born rule)

The complex weights allow destructive interference: configurations with opposite phases cancel. This cancellation is the origin of quantum effects. In the Markov chain, all weights are positive and can only reinforce.

Wick rotation. The Euclidean path integral ($t \rightarrow -i\tau$) replaces e^{iS_M} with e^{-S_E} , $S_E \geq 0$. The weights become real and positive: the Euclidean path integral *is* a Markov chain (or more precisely, the transfer matrix of a classical statistical-mechanics model). The Boltzmann weight e^{-S_E} is the probability weight of a thermal ensemble, with $\hbar$ playing the role of temperature.

Wick rotation is a computational tool: it converts the oscillatory Minkowski integral into a convergent Euclidean one, making numerical evaluation (lattice Monte Carlo) possible. It does not change the physics. The Fock-layer matrix element is the same; the integration contour in the complex time plane is different.

The classical limit, correctly stated. In the limit $S \gg \hbar$, the phase $e^{iS/\hbar}$ oscillates rapidly and only the stationary-phase configuration (the classical trajectory) contributes to the matrix element. This is a useful approximation inside the Fock layer: it tells you that $\langle x_f | e^{-iHT} | x_i \rangle \approx e^{iS[x_{\text{cl}}]} \cdot F(T)$.

But the classical limit as experienced by the observer—the fact that macroscopic objects have definite positions and momenta—is not explained by stationary phase. It is explained by the Born layer: macroscopic observables have near-delta Born distributions because their eigenvalue degeneracy is exponentially large ($d \sim e^S$) and their variance is suppressed by $1/N$. The observer sees a definite value because the Born distribution has already decided, not because off-classical configurations cancel in a path integral she never computes.

Stationary phase tells you which matrix elements are large. The Born layer tells you what the observer sees. These are different statements in different layers.

8 Wick Rotation: From Oscillation to Sampling

The real-time path integral $\int \mathcal{D}x e^{iS_M[x]}$ has an oscillatory integrand of unit modulus: every configuration contributes with the same absolute weight, and the integral converges (when it converges at all) only by cancellation of phases. The Wick rotation $t \rightarrow -i\tau$ replaces the oscillatory weight e^{iS_M} with a decaying weight e^{-S_E} , converting a delicately cancelling integral into an absolutely convergent one. This is not a change of physics; it is a change of computational strategy inside the Fock layer. But it is the change that makes the path integral a practical tool—and the change that connects quantum mechanics to classical statistical mechanics and to Monte Carlo sampling.

We develop the Wick rotation concretely for the harmonic oscillator, including interactions, before treating quantum field theory.

8.1 The substitution: $t \rightarrow -i\tau$

The Wick rotation is the substitution $t = -i\tau$, or equivalently $\tau = it$, where τ is called the *Euclidean time* (or *imaginary time*). Under this substitution, the time-evolution operator transforms:

$$e^{-iHt} \xrightarrow{t \rightarrow -i\tau} e^{-H\tau}. \quad (8.1)$$

This single replacement changes the character of the evolution completely:

Real time t	Euclidean time τ
e^{-iHt} : unitary (rotates)	$e^{-H\tau}$: positive, contracting (decays)
$e^{-iE_n t}$: oscillates on the unit circle	$e^{-E_n \tau}$: decays exponentially
$\ e^{-iHt} \psi\rangle\ = 1$ for all t	$\ e^{-H\tau} \psi\rangle\ \leq 1$, decaying
All eigenstates persist equally	Excited states suppressed; ground state dominates

In real time, the engine rotates the hidden state: all energy eigenstates persist with equal amplitude, each acquiring a phase $e^{-iE_n t}$. In Euclidean time, the engine contracts the hidden state: each eigenstate is suppressed by $e^{-E_n \tau}$, and at large τ only the ground state survives. Real time preserves the state; Euclidean time projects onto the ground state.

8.2 The harmonic oscillator in both times

The harmonic oscillator propagator was computed exactly in section 7:

$$K(x_f, T; x_i, 0) = \sqrt{\frac{\omega}{2\pi i \sin \omega T}} \exp\left(\frac{i\omega}{2 \sin \omega T} [(x_i^2 + x_f^2) \cos \omega T - 2x_i x_f]\right). \quad (8.2)$$

Now set $T = -i\tau$. Under $T \rightarrow -i\tau$:

$$\sin \omega T \rightarrow \sin(-i\omega\tau) = -i \sinh \omega\tau, \quad \cos \omega T \rightarrow \cos(-i\omega\tau) = \cosh \omega\tau. \quad (8.3)$$

Substituting:

$$\boxed{K_E(x_f, \tau; x_i, 0) = \sqrt{\frac{\omega}{2\pi \sinh \omega\tau}} \exp\left(-\frac{\omega}{2 \sinh \omega\tau} [(x_i^2 + x_f^2) \cosh \omega\tau - 2x_i x_f]\right)}. \quad (8.4)$$

Compare the two propagators term by term:

Real-time $K(x_f, T; x_i, 0)$	Euclidean $K_E(x_f, \tau; x_i, 0)$
Prefactor: $\sqrt{\omega/(2\pi i \sin \omega T)}$ complex, oscillatory	Prefactor: $\sqrt{\omega/(2\pi \sinh \omega \tau)}$ real, positive
Exponent: $i\omega[\dots]/(2 \sin \omega T)$ purely imaginary	Exponent: $-\omega[\dots]/(2 \sinh \omega \tau)$ purely real, negative
$ K $ oscillates; no decay	K_E decays exponentially in $(x_i^2 + x_f^2)$
$\sin \omega T = 0$ at $T = n\pi/\omega$: caustics	$\sinh \omega \tau > 0$ for $\tau > 0$: no singularities

The real-time propagator is complex, oscillatory, and singular at focal points. The Euclidean propagator is real, positive, smooth, and decaying—a Gaussian in (x_i, x_f) .

The spectral representation. Both propagators have the same spectral decomposition, differing only in the time dependence:

$$K(x_f, T; x_i, 0) = \sum_{n=0}^{\infty} \psi_n(x_f) \psi_n^*(x_i) e^{-iE_n T}, \quad (8.5)$$

$$K_E(x_f, \tau; x_i, 0) = \sum_{n=0}^{\infty} \psi_n(x_f) \psi_n^*(x_i) e^{-E_n \tau}. \quad (8.6)$$

In real time, each term oscillates at frequency E_n ; the sum oscillates forever without decay. In Euclidean time, each term decays at rate E_n ; at large τ , only the ground state ($n = 0$, slowest decay) survives:

$$K_E(x_f, \tau; x_i, 0) \xrightarrow{\tau \rightarrow \infty} \psi_0(x_f) \psi_0^*(x_i) e^{-E_0 \tau}. \quad (8.7)$$

This is the ground-state projection: the Euclidean propagator at large τ gives the ground-state wavefunction and the ground-state energy. The energy gap $\Delta E = E_1 - E_0$ controls the rate at which excited states die out: $K_E(\tau) \approx e^{-E_0 \tau} [1 + O(e^{-\Delta E \tau})]$.

8.3 The Euclidean path integral: convergent and positive

Under $t \rightarrow -i\tau$, the path integral weight transforms. The Minkowski action for the harmonic oscillator is $S_M = \int dt [\dot{x}^2/2 - \omega^2 x^2/2]$. Setting $t = -i\tau$: $dt = -id\tau$, $\dot{x} = dx/dt = i dx/d\tau \equiv i\dot{x}_E$, so $\dot{x}^2 = -\dot{x}_E^2$. The Minkowski action becomes:

$$iS_M = i \int dt \left[\frac{\dot{x}^2}{2} - \frac{\omega^2 x^2}{2} \right] = - \int d\tau \left[\frac{\dot{x}_E^2}{2} + \frac{\omega^2 x^2}{2} \right] = -S_E, \quad (8.8)$$

where:

$$S_E[x_E] = \int_0^\beta d\tau \left[\frac{1}{2} \dot{x}_E^2 + \frac{1}{2} \omega^2 x^2 \right]. \quad (8.9)$$

The Euclidean action has kinetic energy *plus* potential energy—both terms positive. The path integral weight becomes:

$$e^{iS_M} \rightarrow e^{-S_E}. \quad (8.10)$$

This changes everything about convergence:

Minkowski weight e^{iS_M}	Euclidean weight e^{-S_E}
Complex, $ e^{iS_M} = 1$	Real, $0 < e^{-S_E} \leq 1$
Every configuration contributes equally in modulus	Large-action configurations exponentially suppressed
Convergence by phase cancellation only	Absolute convergence: $\int \mathcal{D}x_E e^{-S_E} < \infty$
Gaussian integral defined by analytic continuation	Gaussian integral defined by standard methods
Not directly sampleable	Sampleable by Monte Carlo

Have we already Wick-rotated? Yes—implicitly. In section 7, the harmonic oscillator path integral was evaluated using the Gaussian integral formula $\int \mathcal{D}\xi e^{iS_{\text{HO}}[\xi]} = F(T) = \sqrt{\omega/(2\pi i \sin \omega T)}$. But the fluctuation integrand $e^{iS_{\text{HO}}[\xi]}$ is oscillatory with unit modulus—it does not converge absolutely. The formula $F(T)$ is obtained by computing the convergent Euclidean integral $\int \mathcal{D}\xi_E e^{-S_E[\xi_E]} = \sqrt{\omega/(2\pi \sinh \omega T)}$ and then analytically continuing $\tau \rightarrow iT$ (i.e., $\sinh \omega \tau \rightarrow i \sin \omega T$).

The real-time Gaussian integral formula is the analytic continuation of the Euclidean one. The Euclidean integral is the definition; the real-time integral is the continuation. Whenever you write $\int dx e^{iax^2/2} = \sqrt{2\pi i/a}$ for real a , you are using the analytic continuation of $\int dx e^{-ax^2/2} = \sqrt{2\pi/a}$ from positive a to imaginary a . The real-time path integral inherits its meaning from the Euclidean one.

This is not a technicality. For the free and harmonic oscillator, the analytic continuation is straightforward because both sides are Gaussian. For the interacting oscillator, the real-time path integral is not Gaussian and cannot be evaluated in closed form; the Euclidean path integral is still well-defined (positive weights, convergent integral) and is the natural starting point for perturbation theory and for numerical computation.

8.4 The interacting case: Euclidean perturbation theory

Add the anharmonic interaction $\lambda x^4/4!$. The Euclidean action becomes:

$$S_E[x_E] = \int_0^\beta d\tau \left[\frac{1}{2} \dot{x}_E^2 + \frac{1}{2} \omega^2 x^2 + \frac{\lambda}{4!} x^4 \right]. \quad (8.11)$$

For $\lambda > 0$, all three terms are non-negative: the Euclidean action is bounded below, the integral $\int \mathcal{D}x_E e^{-S_E}$ converges absolutely, and the weight e^{-S_E} is a well-defined positive measure on the space of configurations.

The perturbative expansion proceeds as in the Minkowski case: Taylor-expand $e^{-\frac{\lambda}{4!} \int x^4 d\tau}$ in powers of λ and evaluate by Wick's theorem. The Wick contractions now use the *Euclidean* Green's function:

$$G_E(\tau_1, \tau_2) = \langle x_E(\tau_1) x_E(\tau_2) \rangle_{S_{E,0}} = \frac{1}{2\omega} e^{-\omega|\tau_1 - \tau_2|}, \quad (8.12)$$

which decays exponentially in $|\tau_1 - \tau_2|$ (contrast the real-time Green's function, which oscillates). The Feynman diagrams are the same—the same vertices, the same combinatorics, the same symmetry factors—but every propagator line is a decaying exponential instead of an oscillating phase. The Euclidean diagrams converge better; ultraviolet divergences are the same (they come from short-distance singularities $\tau_1 \rightarrow \tau_2$), but infrared behaviour is tamer.

Phase becomes noise: the deep correspondence. The Wick rotation exchanges two fundamentally different mechanisms for selecting the dominant configurations.

In real time, the path integral weight $e^{iS_M[x]}$ is a pure phase: every configuration contributes with the same absolute weight $|e^{iS_M}| = 1$. The integral converges by *destructive interference*: configurations with large action have wildly oscillating phases and cancel each other when summed. Only configurations near the stationary-phase point (the classical trajectory, where $\delta S_M = 0$) contribute coherently. The “noise” in the real-time path integral is complex-valued oscillation: the sum over configurations is a sum of unit-modulus complex numbers pointing in all directions, with only a narrow band of configurations constructively reinforcing.

In Euclidean time, the weight $e^{-S_E[x_E]}$ is a real positive number: it is a probability weight, not a phase. Large-action configurations are exponentially suppressed: $e^{-S_E} \rightarrow 0$ as $S_E \rightarrow \infty$. The integral converges by *Gaussian suppression*: the weight assigns high probability to low-action configurations and negligible probability to high-action ones. The “noise” is real-valued Gaussian fluctuation: the configurations sampled from e^{-S_E} cluster around the minimum-action configuration with Gaussian spread, exactly as thermal fluctuations cluster around the energy minimum in a classical system at temperature $T = \hbar$.

The correspondence is precise:

	Minkowski (e^{iS_M})	Euclidean (e^{-S_E})
Weight	Phase on unit circle: $ w = 1$	Positive real: $0 < w \leq 1$
Large S	Rapid oscillation $\rightarrow$ cancellation	Exponential suppression $\rightarrow$ negligible
Dominant configs	Near $\delta S_M = 0$ (stationary phase)	Near $\min S_E$ (Boltzmann peak)
Selection mechanism	Destructive interference	Gaussian decay
“Noise” character	Complex oscillation (interference)	Real Gaussian fluctuation (thermal)
Mathematical tool	Stationary-phase approximation	Saddle-point / Laplace method
Numerical access	Not directly sampleable	Monte Carlo sampling

Both sides select the same configurations (those with small action), but the mechanism differs: on the Minkowski side, it is the interference of complex amplitudes; on the Euclidean side, it is the suppression of a Gaussian probability distribution. Quantum interference in real time becomes classical thermal fluctuation in imaginary time. This is the deepest content of the Wick rotation.

8.5 The Euclidean path integral as classical statistical mechanics

The Euclidean path integral with periodic boundary conditions $x_E(0) = x_E(\beta)$ is the thermal partition function:

$$\mathcal{Z}(\beta) = \text{Tr } e^{-\beta H} = \int dx \, K_E(x, \beta; x, 0) = \int_{\text{periodic}} \mathcal{D}x_E \, e^{-S_E[x_E]}. \quad (8.13)$$

The inverse temperature $\beta = 1/k_B T$ plays the role of the length of the Euclidean time interval, and $\hbar$ (set to 1 throughout) sets the scale.

This integral is *exactly* a classical statistical mechanics partition function. To see this, discretise the Euclidean time axis into N_τ steps with spacing $\epsilon = \beta/N_\tau$:

$$\mathcal{Z}(\beta) = \int \prod_{k=0}^{N_\tau-1} dx_k \, \exp\left(-\epsilon \sum_{k=0}^{N_\tau-1} \left[\frac{1}{2} \left(\frac{x_{k+1} - x_k}{\epsilon} \right)^2 + V(x_k) \right] \right), \quad (8.14)$$

where $x_{N_\tau} = x_0$ (periodic). This is the partition function of a one-dimensional classical chain:

QM in Euclidean time	Classical 1D chain
Position x_k at imaginary time τ_k	Displacement of site k in the chain
Kinetic term $(x_{k+1} - x_k)^2/2\epsilon$	Nearest-neighbour spring coupling
Potential $V(x_k)$	On-site potential energy
Euclidean time extent β	Number of sites $\times$ lattice spacing
$\hbar$	Temperature of the classical chain
Quantum ground state ($\beta \rightarrow \infty$)	Zero-temperature limit of the chain

The quantum harmonic oscillator in Euclidean time *is* a classical chain of harmonic springs. Adding the interaction $\lambda x^4/4!$ adds an anharmonic on-site potential to the chain. The quantum interacting oscillator in Euclidean time is an anharmonic classical chain.

8.6 Quantum Monte Carlo: sampling the Fock layer

The discretised Euclidean partition function (21.10) is a finite-dimensional integral with a positive weight $e^{-S_E} \geq 0$. It can be evaluated numerically by *Monte Carlo sampling*: generate configurations $\{x_0, x_1, \dots, x_{N_\tau-1}\}$ from the probability distribution $P[\{x_k\}] \propto e^{-S_E[\{x_k\}]}$ using the Metropolis algorithm or hybrid Monte Carlo, and estimate expectation values as sample averages:

$$\langle f \rangle = \frac{1}{\mathcal{Z}} \int \prod_k dx_k f(\{x_k\}) e^{-S_E} \approx \frac{1}{N_{\text{samples}}} \sum_{s=1}^{N_{\text{samples}}} f(\{x_k^{(s)}\}). \quad (8.15)$$

What quantum Monte Carlo computes. QMC estimates Euclidean correlation functions, from which Born-layer observables are extracted:

Energy levels. The Euclidean two-point function $C(\tau) = \langle x_E(\tau) x_E(0) \rangle_E$, measured by Monte Carlo, decays as:

$$C(\tau) = \sum_n |\langle n | \hat{x} | 0 \rangle|^2 e^{-(E_n - E_0)\tau} \xrightarrow{\tau \rightarrow \infty} |\langle 1 | \hat{x} | 0 \rangle|^2 e^{-(E_1 - E_0)\tau}. \quad (8.16)$$

The energy gap $E_1 - E_0$ is extracted from the exponential decay rate at large τ . Higher energy gaps come from fitting the correlation function at shorter τ or from correlator matrices.

Ground-state wavefunction. The histogram of x_0 values sampled from $P[\{x_k\}] \propto e^{-S_E}$ at large β converges to $|\psi_0(x)|^2$: the Born probability distribution for position in the ground state.

Thermodynamic quantities. Energy $\langle H \rangle = -\partial \ln \mathcal{Z} / \partial \beta$, specific heat, free energy—all extracted from $\mathcal{Z}(\beta)$ and its derivatives.

The sampled configurations are not paths. Each Monte Carlo configuration $\{x_0, x_1, \dots, x_{N_\tau-1}\}$ is a set of bookkeeping eigenvalues at discrete imaginary-time points. The sampling algorithm generates these configurations with frequency proportional to e^{-S_E} . It is tempting to call each configuration a “path in imaginary time” and to narrate the Monte Carlo history as “the system exploring different trajectories.” This is the same conflation as in the real-time case: the configurations are basis labels in a matrix computation, not Born-layer actualities. The system does not move through imaginary time. The Euclidean “time” τ is a bookkeeping parameter that indexes the classical chain, not a physical time in which events occur.

QMC samples the Fock layer’s computational structure, not the Born layer’s rendered output. The Born-layer observables (energy levels, wavefunctions, thermodynamic quantities) are extracted from the sampled correlators by the same bridges described in section 7.8: spectral decomposition

for masses, histograms for Born distributions, thermodynamic derivatives for macroscopic observables.

8.7 The sign problem: when Monte Carlo fails

The Euclidean path integral is sampleable because $e^{-S_E} \geq 0$ for a single real scalar field with $\lambda \geq 0$. But this positivity fails in several physically important cases:

Fermions. The Grassmann integration over fermionic fields produces a determinant $\det M$ that can be negative. The effective weight $e^{-S_E^{\text{bos}}} \det M$ is not positive definite, and the Monte Carlo interpretation of the Euclidean integral as a probability distribution breaks down. In lattice QCD, clever algorithms (hybrid Monte Carlo with the determinant absorbed into the measure) handle this for zero chemical potential, but at finite baryon density $\mu_B \neq 0$, the determinant becomes complex and the sign problem is severe.

Finite chemical potential. $S_E \rightarrow S_E - \mu \int d\tau \hat{N}$ introduces a complex phase, and the weight is no longer positive.

Frustrated spin systems and some condensed matter models. Geometrical frustration can produce negative Boltzmann weights in the effective classical chain.

The sign problem is a Fock-layer computational obstacle. It is a problem with the Euclidean path-integral *method*, not with the physics. The Hamiltonian is perfectly well-defined: H is a Hermitian operator, its eigenvalues are real, and its eigenstates exist regardless of whether the Euclidean path integral can be sampled. The sign problem is the statement that one particular Fock-layer computational tool (Monte Carlo sampling of e^{-S_E}) fails for certain Hamiltonians.

This is where the Hamiltonian methods—exact diagonalisation, DMRG, tensor networks, Hamiltonian truncation, quantum simulation—become essential. They work directly with H and e^{-iHt} (or $e^{-H\tau}$) without passing through a path integral. They do not have a sign problem, because they never construct a probability measure to sample from—they compute matrix elements directly.

The sign problem is a strong argument for keeping the Hamiltonian at the centre of the formalism, as the Fock-Born framework does. The path integral is a powerful tool when it works (positive Euclidean weights), but the Hamiltonian is always there, and it always works.

8.8 Summary: the Fock-Born perspective on Wick rotation

Wick rotation is a change of computational strategy inside the Fock layer. It replaces one method of computing Fock-layer matrix elements (oscillatory real-time integral, convergent only by phase cancellation) with another (decaying Euclidean integral, absolutely convergent, sampleable by Monte Carlo). The matrix elements being computed are the same; the contour of integration in the complex time plane is different.

	Real time	Euclidean time
Engine operation	e^{-iHt} : rotation	$e^{-H\tau}$: contraction
Weight of each configuration	e^{iS_M} : oscillatory	e^{-S_E} : decaying
Convergence	Phase cancellation	Absolute
Computation method	Analytic continuation, perturbation theory	Monte Carlo (when sign-free), perturbation theory
Physical time evolution	Yes: $h_t = e^{-iHt}h_0$	No: τ is a computational parameter
Ground-state access	All eigenstates persist equally	Ground state dominates at large τ
Layer	Fock	Fock

Both the real-time and Euclidean path integrals are Fock-layer tools. Neither produces Born-layer observables directly. The bridge to Born-layer observables—spectral decomposition for masses, LSZ for cross sections, histograms for Born distributions—is the same in both cases. Wick rotation changes the computational tool; the Fock-Born architecture stays fixed.

The real-time path integral is the one that corresponds to the physical Schrödinger evolution $h_t = e^{-iHt}h_0$. But the Euclidean path integral is the one that is mathematically well-defined (absolutely convergent) and computationally tractable (sampleable by Monte Carlo). In practice, most non-perturbative path-integral computations—lattice QCD, lattice condensed-matter models, quantum Monte Carlo for molecular systems—are done in Euclidean time, with real-time physics recovered by analytic continuation of the results.

The Hamiltonian, however, operates in real time by definition: $h_t = e^{-iHt}h_0$ is the engine. It does not need Wick rotation. It does not need a positive measure. It does not have a sign problem. The Hamiltonian is always available as the primary tool, and the path integral—real-time or Euclidean—is a calculational device derived from it.

8.9 Wick rotation in the Fock-Born framework

The Wick rotation acts on the Fock layer. It does not act on the Born layer. This distinction is essential and is often blurred in textbooks.

What happens to the Fock layer. The Fock layer is the engine: $h_t = e^{-iHt}h_0$. Wick rotation replaces the engine’s operation:

$$\underbrace{e^{-iHt}}_{\text{rotation (unitary)}} \xrightarrow{t \rightarrow -i\tau} \underbrace{e^{-H\tau}}_{\text{contraction (non-unitary)}}. \quad (8.17)$$

In real time: the engine rotates h_t within Hilbert space, preserving its norm $\|h_t\| = 1$. Every energy eigenstate persists with equal amplitude; only the phases change. The engine is unitary, reversible, and norm-preserving.

In Euclidean time: the engine contracts h_t toward the ground state. Each energy eigenstate is multiplied by $e^{-E_n\tau}$; excited states ($E_n > E_0$) are exponentially suppressed. At large τ , only the ground state survives: $e^{-H\tau}h_0 \approx \langle E_0|h_0 \rangle e^{-E_0\tau} |E_0\rangle$. The engine is non-unitary, irreversible, and norm-destroying. (The state must be renormalised to extract the ground-state projection.)

In ML terms: real-time evolution is like an RNN with unitary weight matrix (all eigenvalues on the unit circle, no information loss). Euclidean evolution is like an RNN with eigenvalues inside the unit circle: the hidden state decays toward the dominant eigenvector (power iteration). The Euclidean engine is a power-iteration algorithm for finding the ground state.

What happens to the Born layer. The Born layer is the interface: the observer submits an observable $\hat{\mathcal{O}}$ and receives an eigenvalue λ with probability $P(\lambda) = \|P_\lambda h_t\|^2$.

Under Wick rotation: the Born layer *does not operate*. Euclidean time τ is a computational parameter, not a physical time. There is no observer at Euclidean time τ submitting observables and receiving outcomes. The Born protocol—probability, sampling, trimming—requires real time, because it describes the interaction between the engine and a physical observer, and physical observers exist in real time.

The Euclidean path integral produces *Fock-layer intermediaries* (correlation functions, partition functions, propagators). These are converted to *Born-layer observables* (energy levels, cross sections, Born distributions) by the bridges of section 7.8: spectral decomposition, LSZ, thermodynamic derivatives. The bridges operate after the Euclidean computation is complete, converting the results back to real-time predictions.

Layer	Real time	Euclidean time
Fock	$h_t = e^{-iHt}h_0$: rotation (norm-preserving)	$h_\tau \propto e^{-H\tau}h_0$: contraction (norm-destroying)
Born	Operational: observer submits $\hat{\mathcal{O}}$, receives λ , Born probability $P(\lambda) = \ P_\lambda h_t\ ^2$	Not operational: no observer at imaginary time; τ is a computational parameter
Output	Born-layer eigenvalue (a measurement result)	Fock-layer matrix element (a number inside the engine)

The contour rotation is legitimate because H is bounded below. Why can we rotate the time contour from the real axis to the imaginary axis without encountering singularities? Because H has a ground state: $E_n \geq E_0$ for all n . The matrix element $\langle f | e^{-iHt} | i \rangle = \sum_n \langle f | E_n \rangle \langle E_n | i \rangle e^{-iE_n t}$ is a sum of terms $e^{-iE_n t}$. In the lower-half complex t -plane ($\text{Im } t < 0$, i.e. $t = -i\tau + t_R$ with $\tau > 0$), each term becomes $e^{-E_n\tau} e^{-iE_n t_R}$, which decays as τ increases because $E_n \geq E_0 > -\infty$. No term blows up; no pole is crossed. The contour rotation from $t \in \mathbb{R}$ to $t = -i\tau$ (a 90° rotation into the lower-half plane) is an analytic continuation in a region free of singularities.

If H were not bounded below (negative-energy states extending to $-\infty$), the terms $e^{-E_n\tau}$ with $E_n < 0$ would *grow* as $\tau \rightarrow \infty$, and the Euclidean path integral would diverge. The Wick rotation works precisely because the Hamiltonian has a ground state.

Euclidean correlators encode Born-layer physics. Although the Born layer does not operate at Euclidean time, the Euclidean correlators encode all the information needed to reconstruct Born-layer predictions at real time. The key example:

The Euclidean two-point function $C_E(\tau) = \langle x_E(\tau) x_E(0) \rangle_E$ decays as:

$$C_E(\tau) = \sum_n |\langle n | \hat{x} | 0 \rangle|^2 e^{-(E_n - E_0)\tau}. \quad (8.18)$$

At large τ : $C_E(\tau) \approx |\langle 1 | \hat{x} | 0 \rangle|^2 e^{-\Delta E \tau}$, where $\Delta E = E_1 - E_0$ is the energy gap. The energy gap is a Born-layer quantity: it is what a spectrometer reports when it measures the photon emitted in the transition $|1\rangle \rightarrow |0\rangle$.

The Euclidean correlator, computed by Monte Carlo sampling of e^{-S_E} (Gaussian-noise configurations in imaginary time), encodes the energy gap (and all higher gaps) as exponential decay rates. The Gaussian noise of the Euclidean path integral carries the same spectral information as the quantum interference of the Minkowski path integral—but in a form that is computationally accessible.

This is the practical payoff of the phase-noise correspondence: quantum observables (energy levels, decay rates, scattering cross sections) can be extracted from classical noise (Monte Carlo samples of e^{-S_E}), because the two are related by analytic continuation in the complex time plane. The Euclidean noise is a computational representation of the Minkowski interference, and the Born-layer observables encoded in one are the same as those encoded in the other.

9 Degenerate Observables and the Projected Superposition

In the previous sections, every observable had distinct eigenvalues: each eigenvalue belonged to exactly one eigenvector, and the Born rule gave $P(\lambda) = |\langle \lambda | h_t \rangle|^2$. The Bohr trim set $h_{t+} = |\lambda\rangle$: a complete reset to a single eigenvector.

This was the simplest case. We now confront the general case, which is far richer and far more important. Most observables in nature have *degenerate* eigenvalues: eigenvalues shared by many eigenvectors. The degenerate Born rule is the central mechanism of quantum mechanics—more central than interference, more central than entanglement, more central than any specific Hamiltonian. It is the mechanism by which quantum amplitudes become classical probabilities, by which temperature emerges, and by which the classical world appears from the Fock-layer engine.

9.1 The identity observable: the starting point

The identity operator $\hat{O} = \mathbf{1}$ is Hermitian, with one eigenvalue $\lambda = 1$ and eigensubspace $\mathcal{H}_1 = \mathcal{H}$: every vector is an eigenvector. Degeneracy $d_1 = \dim \mathcal{H}$: maximal.

Apply the Born protocol:

1. **Born probability.** $P(1) = \|P_1 h_t\|^2 = \|\mathbf{1} h_t\|^2 = \|h_t\|^2 = 1$. Certainty.
2. **Born sampling.** The sample is $\lambda = 1$ with probability 1. Deterministic.
3. **Bohr trimming.** $h_{t+} = P_1 h_t / \|P_1 h_t\| = h_t / 1 = h_t$. The state is unchanged.

The identity asks no question, distinguishes no states, learns nothing, trims nothing. It is the measurement that does not measure.

The identity is not a pathology. It is the *starting point* of the entire Born hierarchy. Every other observable is a *refinement* of the identity: it splits the single eigensubspace $\mathcal{H}$ into smaller subspaces, each with a distinct eigenvalue. The more the eigensubspace is split, the more the measurement learns and the more the Bohr trim removes.

9.2 Partial lifting of degeneracy: a concrete example

Consider $\mathcal{H} = \mathbb{C}^4$ with basis $\{|1\rangle, |2\rangle, |3\rangle, |4\rangle\}$ and hidden state $h_t = \alpha|1\rangle + \beta|2\rangle + \gamma|3\rangle + \delta|4\rangle$, $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$.

Four observables with decreasing degeneracy:

Observable A: the identity ($d = 4$). One eigenvalue $\lambda = 1$, degeneracy 4. Born probability: $P(1) = 1$. Projected superposition: h_t itself. Trim: nothing. The Born distribution is a delta at $\lambda = 1$.

Observable B: two eigenvalues ($d = 2, 2$). $\hat{O}_B = \text{diag}(+1, +1, -1, -1)$. Eigenvalue $+1$: eigensubspace $\text{span}\{|1\rangle, |2\rangle\}$, degeneracy 2. Eigenvalue -1 : eigensubspace $\text{span}\{|3\rangle, |4\rangle\}$, degeneracy 2.

Born probabilities:

$$P(+1) = |\alpha|^2 + |\beta|^2, \tag{9.1}$$

$$P(-1) = |\gamma|^2 + |\delta|^2. \tag{9.2}$$

The probabilities are *sums* of squared amplitudes—the OR rule: the probability of $+1$ is the sum of the probabilities for $|1\rangle$ and $|2\rangle$, because the degeneracy means the observable cannot

distinguish them. No interference between $|1\rangle$ and $|2\rangle$ contributes to $P(+1)$: the cross terms $\alpha^*\beta + \beta^*\alpha$ are absent, because $|1\rangle$ and $|2\rangle$ are orthogonal within the eigensubspace.

If outcome $+1$ is observed, the Bohr trim gives:

$$h_{t+} = \frac{\alpha |1\rangle + \beta |2\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}}. \quad (9.3)$$

The projected superposition: everything in $\text{span}\{|3\rangle, |4\rangle\}$ is discarded; the superposition within $\text{span}\{|1\rangle, |2\rangle\}$ survives with its original amplitudes and phases. The measurement learns “ $+1$ vs -1 ” but does not learn “ $|1\rangle$ vs $|2\rangle$.”

Observable C: three eigenvalues ($d = 2, 1, 1$). $\hat{\mathcal{O}}_C = \text{diag}(0, 0, +1, -1)$. Eigenvalue 0 has degeneracy 2; eigenvalues ± 1 have degeneracy 1.

If outcome 0 is observed: $h_{t+} = (\alpha |1\rangle + \beta |2\rangle) / \sqrt{|\alpha|^2 + |\beta|^2}$ —a superposition survives. If outcome $+1$ is observed: $h_{t+} = |3\rangle$ —complete collapse.

Observable D: complete measurement ($d = 1, 1, 1, 1$). $\hat{\mathcal{O}}_D = \text{diag}(1, 2, 3, 4)$. Every eigenvalue non-degenerate. Born probabilities: $P(k) = |c_k|^2$ for $k = 1, 2, 3, 4$. Bohr trim: $h_{t+} = |k\rangle$ —complete collapse to a single basis vector. Full information; maximum trimming.

The hierarchy at a glance:

	Degeneracy	Born distribution	Information gained	Superposition after trim
A	(4)	delta	none	full (unchanged)
B	(2, 2)	binary	1 bit	2D survives
C	(2, 1, 1)	ternary	> 1 bit	depends on outcome
D	(1, 1, 1, 1)	quaternary	2 bits	none (single vector)

Degeneracy controls everything: the entropy of the Born distribution, the information gained, and the superposition that survives.

9.3 The general framework

The projected superposition. For an observable $\hat{\mathcal{O}} = \sum_{\lambda} \lambda P_{\lambda}$ with eigensubspace projector $P_{\lambda} = \sum_{k=1}^{d_{\lambda}} |\lambda, k\rangle \langle \lambda, k|$, the projected superposition onto eigenvalue λ is:

$$h_t^{(\lambda)} = P_{\lambda} h_t = \sum_{k=1}^{d_{\lambda}} \langle \lambda, k | h_t \rangle |\lambda, k\rangle. \quad (9.4)$$

This is the piece of h_t compatible with eigenvalue λ . If $d_{\lambda} > 1$, it is a *superposition* of multiple eigenvectors—not a single state.

The Born probability: squared length and the OR rule.

$$P(\lambda) = \|h_t^{(\lambda)}\|^2 = \sum_{k=1}^{d_{\lambda}} |\langle \lambda, k | h_t \rangle|^2. \quad (9.5)$$

The sum over k is the OR rule: the probability of outcome λ is the sum of the individual probabilities for each eigenvector in the eigensubspace. Cross terms vanish by orthogonality. The OR rule is not a new postulate; it is a consequence of the definition of $P(\lambda)$ as a squared length and the orthogonality of the eigenvectors.

The Bohr trim: project, keep, renormalise.

$$h_{t+} = \frac{P_\lambda h_t}{\|P_\lambda h_t\|} = \frac{h_t^{(\lambda)}}{\sqrt{P(\lambda)}}. \quad (9.6)$$

Everything outside $\mathcal{H}_\lambda$ is discarded. Everything inside survives with its original amplitudes and phases, renormalised to unit length. The superposition within the eigensubspace is preserved.

The superposition survives the measurement. This is the most important consequence of degeneracy. When $d_\lambda > 1$, the post-measurement state is not a single eigenvector but a superposition within $\mathcal{H}_\lambda$. The measurement reduces the superposition (trims away the components outside $\mathcal{H}_\lambda$) but does not eliminate it. The surviving superposition carries interference, entanglement, and quantum correlations that the observable's degeneracy prevented the measurement from resolving.

9.4 The roots of degeneracy: why classical observables are degenerate

The Born distribution of an observable is determined by the degeneracy of its eigenvalues. Low degeneracy ($d \sim 1$): the Born distribution can be spread across many eigenvalues, with genuine randomness—the quantum regime. High degeneracy ($d \sim e^{cN}$): the Born distribution concentrates into a near-delta spike—the classical regime.

What makes an observable degenerate? Four mechanisms, each rooted in the mathematical structure of the observable, each producing degeneracy that grows with the size of the system.

Mechanism 1: nonlinear composition. The most important and least appreciated source of degeneracy.

A linear observable—one that is a linear function of the fundamental fields—can have low degeneracy. The spin $\hat{S}_z$ has eigenvalues $\pm\hbar/2$, each non-degenerate. But most observables in physics are *nonlinear composites*: products, squares, or higher functions of the fundamental variables. And nonlinear functions are generically many-to-one.

Example: $r^2 = \sum_{i=1}^N x_i^2$. The value $r^2 = R^2$ is produced by all points on the $(N-1)$ -sphere of radius R : degeneracy $\sim R^{N-1}$. For $N \sim 10^{23}$: $d \sim R^{10^{23}}$. The energy density $\hat{\mathcal{E}}(\mathbf{x}) = \hat{\pi}^2/2 + (\nabla\hat{\phi})^2/2 + m^2\hat{\phi}^2/2$ is quadratic in the fundamental fields—a field-theoretic r^2 .

The general principle: a nonlinear function of N variables has degeneracy that grows at least as a power of N , and exponentially for most functions of practical interest. This is the mechanism that makes $g_{\mu\nu}$ the most classical observable in physics: the metric is a triply nonlinear composite of the fundamental fields, with degeneracies multiplying at each layer.

Mechanism 2: partial access (entangled systems). A composite system $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_E$ observed through $\hat{\mathcal{O}} = \hat{\mathcal{O}}_S \otimes \mathbf{1}_E$ has degeneracy at least $\dim \mathcal{H}_E$ for every system eigenvalue: the identity on the environment means the observable cannot distinguish any environment state.

The Born probability:

$$P(\lambda_S) = \sum_j |\langle \lambda_S, e_j | h_t \rangle|^2 \quad (9.7)$$

is the OR sum over all environment states—classical sum, no interference between different $|e_j\rangle$. This is decoherence: the loss of coherence between environment states, arising entirely from the degeneracy of partial access.

For a macroscopic environment: $\dim \mathcal{H}_E \sim e^{S_E/k_B}$, astronomical degeneracy.

Mechanism 3: statistical and macroscopic averaging. The average observable $\hat{A} = N^{-1} \sum_{k=1}^N \hat{A}_k$ over N commuting degrees of freedom has eigenvalue $\bar{a} = N^{-1} \sum_k a_k$: the

arithmetic mean. Many configurations $(a_1, \dots, a_N)$ produce the same $\bar{a}$. For binary variables: degeneracy $d(\bar{a}) = \binom{N}{N_+} \sim 2^N / \sqrt{N}$ at $\bar{a} = 0$.

The variance:

$$\text{Var}(\hat{A}) = \frac{\overline{\text{Var}(\hat{A}_k)}}{N} \xrightarrow{N \rightarrow \infty} 0. \quad (9.8)$$

The Born distribution becomes a delta function: this is the law of large numbers, restated as a property of the Born output.

This is the mechanism by which temperature, pressure, density, and all macroscopic observables become classical: they are averages over $\sim 10^{23}$ microscopic degrees of freedom. The averaging compresses N eigenvalues into one number, producing exponential degeneracy and $1/N$ variance.

Mechanism 4: finite precision. No physical detector has infinite resolution. A detector with resolution δ measures the binned observable:

$$\hat{X}_\delta = \sum_n x_n P_n, \quad P_n = \int_{x_n - \delta/2}^{x_n + \delta/2} dx |x\rangle \langle x|, \quad (9.9)$$

where P_n projects onto the bin $[x_n - \delta/2, x_n + \delta/2]$. Every eigenvalue x_n acquires degeneracy $\sim \delta \Lambda / \pi$ (the number of momentum modes that fit in a bin of width δ).

Combined with averaging: the physical observable for a macroscopic measurement is the binned average $\hat{A}_\delta$, with degeneracy $\sim e^{S/k_B}$ and Born output deterministic to accuracy $O(e^{-cN})$.

Which observables are classical, and why:

Observable	M1	M2	M3	M4	Character
$\hat{S}_z$ (spin)	–	–	–	weak	Quantum: $d = 1$
$\hat{x}$ (position)	–	moderate	–	on	Quantum: low d
$\hat{\mathcal{E}}$ (energy density)	on	on	–	on	Intermediate
Temperature	on	on	on	on	Classical: $d \sim e^{cN}$
$g_{\mu\nu}$ (metric)	extreme	on	on	on	Maximally classical

The quantum-to-classical transition is not a transition in the physics. It is a transition in the observable: from linear, non-averaged, fully resolved ($d \sim 1$) to nonlinearly composed, partially observed, macroscopically averaged ($d \sim e^{cN}$). The Born rule does not change. The degeneracy changes.

9.5 The Born layer certifies: delta or not

The Fock-Born architecture assigns clear roles. The Fock layer provides the hidden state h_t and the observable $\hat{\mathcal{O}}$. The Born layer does the rest: it computes the probability distribution $P(\lambda) = \|P_\lambda h_t\|^2$ over the eigenvalues, and it *certifies* the character of this distribution.

If the Born distribution is a delta ($P(\lambda_0) = 1$ for some λ_0 , $P(\lambda) = 0$ otherwise): then Born sampling returns λ_0 with certainty, and Bohr trimming leaves h_t unchanged (because h_t already lies entirely within $\mathcal{H}_{\lambda_0}$). The interface is trivially passable. The outcome is a foregone conclusion.

If the Born distribution is broad ($P(\lambda)$ spread across multiple eigenvalues with comparable weights): then Born sampling is genuinely random, and Bohr trimming reshapes h_t drastically. The interface is non-trivial. The outcome is uncertain.

The certification is a Born-layer operation. Determining whether a distribution is delta or broad is itself a computation that happens in the Born layer. The Fock layer provides the ingredients (h_t and $\hat{\mathcal{O}}$); the Born layer computes $P(\lambda)$ and evaluates its concentration. No Fock-layer operation—no amount of Schrödinger evolution, no amount of decoherence, no degree of macroscopic averaging—can substitute for this evaluation.

This is not a pedantic distinction. It separates two logically independent operations:

What the Fock layer does	What the Born layer does
Provides h_t (the hidden state)	Computes $P(\lambda) = \ P_\lambda h_t\ ^2$
Provides $\hat{\mathcal{O}}$ (the observable, with its degeneracy structure)	Certifies whether $P(\lambda)$ is delta, near-delta, or broad
Evolves h_t unitarily (the engine runs)	Samples λ from $P(\lambda)$
May produce h_t such that P is near-delta (through decoherence, averaging, large N)	Trims: $h_{t+} = P_\lambda h_t / \ P_\lambda h_t\ $
Determines <i>which</i> outcome is overwhelmingly probable	Actualises the outcome: transitions possibility into actuality

The Fock layer does the heavy lifting: it determines the Born distribution and, for macroscopic observables, makes it near-delta. The Born layer does the final step: it evaluates the distribution, samples, trims, and actualises.

For macroscopic observables ($d \sim e^{cN}$, variance $\sim 1/N$): the Born layer’s three operations are trivial. Probability: $P(\lambda_0) \approx 1$. Sampling: returns λ_0 with near-certainty. Trimming: $P_{\lambda_0} h_t \approx h_t$. The interface is trivially passable, and the observer sees a definite classical value.

For microscopic observables ($d = 1$, variance $\sim O(1)$): the Born layer’s three operations are non-trivial. The distribution is broad, the sample is random, the trim reshapes the state. The observer sees quantum randomness.

The difference between classical and quantum is not a difference of physics. It is a difference in the Born distribution’s concentration, controlled by the degeneracy of the observable.

Why the Born layer is always needed. Even when $P(\lambda_0) = 1 - 10^{-10^{23}}$, the Born layer is still invoked. A probability distribution, no matter how sharply peaked, is not an outcome. The transition from “ $P(\lambda_0) \approx 1$ ” to “the outcome *is* λ_0 ” is the transition from possibility to actuality—and this transition is the Born layer’s job.

Practically, the distinction is invisible: no experiment can distinguish $P = 1 - 10^{-10^{23}}$ from $P = 1$. Conceptually, the distinction matters for three reasons:

The measurement problem. If the Fock layer alone could actualise outcomes, there would be no measurement problem: the Schrödinger equation would produce definite outcomes for macroscopic observables. The measurement problem exists because the Schrödinger equation preserves superpositions at every N .

The interpretation of probability. The Born rule is not a statement of ignorance (as in classical statistical mechanics). It is a statement about the ontology: h_t is a superposition of possibilities, and the actualisation is genuinely probabilistic (Bell’s theorem). The macroscopic limit does not change the ontology.

The logical structure. The Fock-Born architecture has two layers: engine and interface. If the Born layer were derivable from the Fock layer for macroscopic observables, one layer would be redundant. But the Schrödinger equation preserves superpositions at every N , and the Born rule is an independent postulate at every N . A trivial axiom is still an axiom.

Possibility is not actuality. Narrowing is not selecting. The distribution is not the outcome. And no operation in the Fock layer—however powerful—can cross the bridge from possibility to actuality. That bridge is the Born layer.

9.6 Foreshadowing

The degenerate Born rule developed in this section is the foundation for everything that follows:

Entanglement and decoherence (next section): partial access to a composite system ($\hat{\mathcal{O}}_S \otimes \mathbf{1}_E$) produces degeneracy $\sim \dim \mathcal{H}_E$. The OR sum over the environment is the partial trace. Decoherence is the degenerate Born rule at enormous scale.

Temperature: measuring total energy with multiplicity $W = e^{S/k_B}$ produces the Boltzmann distribution through the OR sum over microstates.

Coherent states and the classical field: the Born distribution for the field observable of a coherent state concentrates with variance $\sim 1/|\alpha|$; the classical EM field is a Born rendering of a highly degenerate field observable.

Condensates: the Born distribution for the $\mathbf{k} = 0$ mode occupation concentrates at $n_0 \sim N_0$ with variance $\sim 1/\sqrt{N_0}$. The order parameter is a Born rendering at the delta extreme.

Confinement: the physical Hilbert space of QCD is the colour-singlet subspace. The projected superposition of any physical state onto a colour-charged basis vector has exactly zero squared length—an exact selection rule from the degeneracy structure of the colour charge observable.

In every case: the degenerate Born rule, the projected superposition, the OR sum, and the squared length. Degeneracy is the single concept that unifies quantum measurement, statistical mechanics, the emergence of classical physics, and the selection rules of the Standard Model.

10 Tensor Product, Entanglement, and Decoherence

The previous section developed the general framework: degeneracy, projected superposition, the OR rule. We now encounter the most physically important *source* of degeneracy: the tensor product structure of composite systems.

When two quantum systems coexist, the joint hidden state is a single vector in the tensor product of the two Hilbert spaces. An observer who measures only one subsystem is submitting an observable that acts as the identity on the other—and the identity, as we have seen, is the maximally degenerate observable. The degeneracy framework applies directly, and it produces entanglement, the density matrix, temperature, and decoherence—all as consequences of the projected superposition mechanism.

10.1 The tensor product: the joint hidden state

For a concrete understanding of the tensor product as vectors and tables, the reader can consult the section on Dirac notation. Here we state the algebraic structure.

The direct product of two kets. Let system A have Hilbert space $\mathcal{H}_A$ with basis $\{|a\rangle\}$ ($a = 1, \dots, n_A$) and system B have $\mathcal{H}_B$ with basis $\{|b\rangle\}$ ($b = 1, \dots, n_B$). If A is in $|\phi_A\rangle = \sum_a \alpha_a |a\rangle$ and B is in $|\phi_B\rangle = \sum_b \beta_b |b\rangle$, the joint state is:

$$|\phi_A\rangle \otimes |\phi_B\rangle = \sum_{a,b} \alpha_a \beta_b |a, b\rangle. \quad (10.1)$$

The joint Hilbert space $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$ has basis $\{|a, b\rangle\}$ and dimension $n_A \times n_B$.

General joint states. A general state in the joint space:

$$h_t = \sum_{a,b} c_{ab} |a, b\rangle, \quad \sum_{a,b} |c_{ab}|^2 = 1. \quad (10.2)$$

There is no “state of A ” and separate “state of B ”; there is only the joint hidden state h_t . A *product state* has $c_{ab} = \alpha_a \beta_b$; an *entangled state* does not factorise.

The direct product of operators. $O_A \otimes O_B$ acts on each factor independently:

$$(O_A \otimes O_B) \sum_{a,b} c_{ab} |a, b\rangle = \sum_{a,b} c_{ab} (O_A |a\rangle) \otimes (O_B |b\rangle). \quad (10.3)$$

An operator on A alone is $O_A \otimes \mathbf{1}_B$: it acts on A and leaves B unchanged.

10.2 The identity on B : the most important source of degeneracy

This is the central observation that connects the tensor product to the degeneracy framework.

Three levels of observation

Consider the joint space $\mathcal{H}_A \otimes \mathcal{H}_B$. Three observables, ordered by decreasing degeneracy:

Observable	Degeneracy per eigenvalue	Distinct eigenvalues	What it learns
$\mathbf{1}_A \otimes \mathbf{1}_B$	$n_A n_B$	1	Nothing
$\hat{O}_A \otimes \mathbf{1}_B$	n_B	n_A	A ’s eigenvalue; nothing about B
$\hat{O}_A \otimes \hat{O}_B$	1	$n_A n_B$	Both A and B

The progression mirrors the $A \rightarrow D$ hierarchy of the previous section, but now the degeneracy has a physically meaningful source: the unobserved subsystem.

The restricted observable $\hat{\mathcal{O}}_A \otimes \mathbf{1}_B$

An observer with access to A but not B measures an A -observable $\hat{\mathcal{O}}_A$ with eigenstates $|a_o\rangle$ and eigenvalues a_o . The operator submitted to the Born layer is:

$$\hat{\mathcal{O}}_A \otimes \mathbf{1}_B. \quad (10.4)$$

Its eigenvalue a_o has eigenvectors $\{|a_o, b\rangle\}$ for every b :

$$(\hat{\mathcal{O}}_A \otimes \mathbf{1}_B) |a_o, b\rangle = a_o |a_o, b\rangle \quad \text{for all } b \in \{1, \dots, n_B\}. \quad (10.5)$$

The degeneracy is $d_{a_o} = n_B$: the identity $\mathbf{1}_B$ cannot distinguish any states of B , so every $|b\rangle$ is an eigenvector with the same eigenvalue. The unobserved subsystem B is the source of the degeneracy.

10.3 Measuring A alone: the Born-Bohr rule applied

We apply the three-step Born-Bohr rule from the previous section to $\hat{\mathcal{O}}_A \otimes \mathbf{1}_B$.

Step 1: project. The projected superposition of h_t onto eigenvalue a_o :

$$h_t^{(a_o)} = \sum_{b=1}^{n_B} c(a_o, b) |a_o, b\rangle. \quad (10.6)$$

This is the piece of h_t with A -eigenvalue a_o , retaining all B -amplitudes with their original magnitudes and phases.

Step 2: squared length (the OR rule over B).

$$P(a_o) = \|h_t^{(a_o)}\|^2 = \sum_{b=1}^{n_B} |c(a_o, b)|^2. \quad (10.7)$$

Cross terms vanish by $\langle b|b'\rangle = \delta_{bb'}$. The probability of a_o is the sum of the individual probabilities over all states of the unobserved subsystem B —the classical OR rule. This is a marginalisation: $P(a_o) = \sum_b P(a_o, b)$.

Step 3: trim and renormalize. If the observer obtains a_o :

$$h_{t+} = \frac{\sum_b c(a_o, b) |a_o, b\rangle}{\sqrt{\sum_b |c(a_o, b)|^2}}. \quad (10.8)$$

The A -part is definite (a_o). The B -part is still a superposition with its original coefficients intact. The collapse has resolved A and left B untouched.

10.4 Entanglement: when the coefficients do not factorise

Product states: no correlation. If $c_{ab} = \alpha_a \beta_b$:

$$P(a_o) = |\alpha_{a_o}|^2 \underbrace{\sum_b |\beta_b|^2}_{=1} = |\alpha_{a_o}|^2. \quad (10.9)$$

The OR sum over B gives 1, and the Born probability depends only on A . Measuring A reveals nothing about B .

Entangled states: irreducible correlation. An entangled state has $c_{ab} \neq \alpha_a \beta_b$ for any α, β . The Bell state:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|0,0\rangle + |1,1\rangle) \quad (10.10)$$

has $c_{00} = c_{11} = 1/\sqrt{2}$, $c_{01} = c_{10} = 0$. No factorisation exists.

The projected superpositions of the Bell state.

$$h_t^{(0)} = c(0,0)|0,0\rangle + c(0,1)|0,1\rangle = \frac{1}{\sqrt{2}}|0,0\rangle, \quad (10.11)$$

$$h_t^{(1)} = c(1,0)|1,0\rangle + c(1,1)|1,1\rangle = \frac{1}{\sqrt{2}}|1,1\rangle. \quad (10.12)$$

Born probabilities: $P(0) = \frac{1}{2}$, $P(1) = \frac{1}{2}$. If Alice finds 0, the post-measurement state is $|0,0\rangle$: Bob's qubit is also $|0\rangle$. The measurement of A has determined B —not by acting on B , but by trimming the joint projected superposition to a branch where B is definite.

10.5 Bell's theorem: classical values do not exist between measurements

Entanglement tempts the hypothesis that the correlations are pre-established: each particle carries hidden instructions. Bell's theorem rules this out.

The CHSH inequality. Alice measures $\hat{A}$ or $\hat{A}'$ on her qubit; Bob measures $\hat{B}$ or $\hat{B}'$ on his. Each outcome is ± 1 . Define:

$$S = E(A, B) + E(A, B') + E(A', B) - E(A', B'), \quad (10.13)$$

where $E(A, B) = \langle \Phi^+ | (\hat{A} \otimes \hat{B}) | \Phi^+ \rangle$. For the Bell state, $E = \cos \theta$ where θ is the angle between measurement directions.

Classical bound. If each particle carries pre-established values $A(\lambda), B(\lambda) \in \{\pm 1\}$ determined by a hidden variable λ , then for any fixed λ : $A(\lambda)[B(\lambda) + B'(\lambda)] + A'(\lambda)[B(\lambda) - B'(\lambda)] = \pm 2$, because one bracket is ± 2 and the other is 0. Averaging: $|S| \leq 2$.

Quantum violation. Choosing angles $0^\circ, 90^\circ, 45^\circ, -45^\circ$:

$$S = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = 2\sqrt{2} \approx 2.83 > 2. \quad (10.14)$$

Confirmed experimentally with loophole-free precision.

The lesson. The classical variables $A(\lambda), B(\lambda)$ do not exist before measurement. The outcomes are produced by the Born rendering layer at the moment of measurement, not read off from pre-existing hidden variables. Between measurements, there are no classical values—only h_t rotating in Fock space.

10.6 Spooky action is not spooky: it is just h_t

When Alice measures and obtains 0:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|0,0\rangle + |1,1\rangle) \xrightarrow{\text{Alice finds 0}} |0,0\rangle. \quad (10.15)$$

Bob's qubit is now $|0\rangle$, even if Bob is far away.

The joint hidden state $h_t = |\Phi^+\rangle$ is a single vector in the game engine. It has no spatial location. When the Born layer samples 0 and the Bohr layer trims h_t to $|0,0\rangle$, a single vector in the

engine's memory has been updated. This update is instantaneous in the game engine, because the game engine does not live in the rendered spacetime—it runs the rendered spacetime.

No information travels faster than light. The Born layer samples 0 or 1 randomly; Alice cannot control which. The no-signalling theorem is a theorem about the Born rendering layer: the randomness of the Born output prevents exploitation of the correlation for communication.

10.7 The density matrix from the degeneracy framework

The density matrix is not a new postulate. It is a packaging of the projected superpositions for a restricted observer.

The expectation value leads to ρ_A

For a general A -observable $\hat{O}_A$ with matrix elements $(\hat{O}_A)_{a,a'}$:

$$\begin{aligned}\langle \hat{O}_A \rangle &= \langle h_t | (\hat{O}_A \otimes \mathbf{1}_B) | h_t \rangle \\ &= \sum_{a,a'} \underbrace{\left(\sum_b c(a,b) c^*(a',b) \right)}_{(\rho_A)_{a,a'}} (\hat{O}_A)_{a',a}.\end{aligned}\tag{10.16}$$

The object that naturally appears is the *reduced density matrix*:

$$\boxed{(\rho_A)_{a,a'} = \sum_b c(a,b) c^*(a',b) = \text{Tr}_B[|h_t\rangle \langle h_t|]}.\tag{10.17}$$

The partial trace sums over B 's basis states—the same OR sum over the degenerate index b that produced the Born probabilities.

What ρ_A encodes

Diagonal entries: $(\rho_A)_{a,a} = \sum_b |c(a,b)|^2 = P(a)$. The squared lengths of the projected superpositions—the Born probabilities from the OR rule over B .

Off-diagonal entries: $(\rho_A)_{a,a'} = \sum_b c(a,b) c^*(a',b)$ for $a \neq a'$. The *coherences*: cross terms between different projected superpositions. They survive when $c(a,b)$ and $c(a',b)$ have correlated phases across b . If the phases are uncorrelated, the sum over b averages to zero: coherences vanish, ρ_A becomes diagonal.

Purity: $\text{Tr}[\rho_A^2] \leq 1$. Equals 1 if h_t is a product state; equals $1/n_A$ for a maximally entangled state.

The formula: $\langle \hat{O}_A \rangle = \text{Tr}[\rho_A \hat{O}_A]$ for any A -observable.

The Bell state example

For $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|0,0\rangle + |1,1\rangle)$:

$$\rho_A = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \frac{\mathbf{1}}{2}.\tag{10.18}$$

Off-diagonals: $c(0,0)c^*(1,0) + c(0,1)c^*(1,1) = \frac{1}{\sqrt{2}} \cdot 0 + 0 \cdot \frac{1}{\sqrt{2}} = 0$. Alice's reduced state is maximally mixed. All correlations live in the joint state, not in either subsystem.

The density matrix is derived, not postulated

The derivation is a special case of the degeneracy framework:

1. The restricted observable has degeneracy n_B (the identity on B).
2. The projected superpositions retain all B -amplitudes.
3. The OR sum produces the diagonal entries (Born probabilities).
4. The coherences are cross terms between different projected superpositions, suppressed by decoherence when B is large.
5. ρ_A packages everything so that $\langle \hat{\mathcal{O}}_A \rangle = \text{Tr}[\rho_A \hat{\mathcal{O}}_A]$.

The density matrix is a shadow of h_t cast by the restricted viewport—the result of the degenerate Born rule applied to an observer who cannot see B .

10.8 Pure states and mixed states

Pure state: $h_t = |\phi_A\rangle \otimes |\phi_B\rangle$ (product). $\rho_A = |\phi_A\rangle \langle \phi_A|$: rank 1, $\text{Tr}[\rho_A^2] = 1$. A superposition $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ gives:

$$\rho_{\text{pure}} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \quad (10.19)$$

The off-diagonal coherences ($\frac{1}{2}$) encode the superposition. Measuring σ_x : $P(+1) = 1$ (certain)—the coherences produce constructive interference.

Mixed state: A entangled with B (Bell state). $\rho_A = \mathbf{1}/2$: no coherences. Measuring σ_x : $P(+1) = P(-1) = \frac{1}{2}$ —the OR sum over B has killed the interference.

	Pure $ +\rangle \langle + $	Mixed $\mathbf{1}/2$
Density matrix	$\frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	$\frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Coherences	Present	Absent (OR sum killed them)
σ_z measurement	$P(\pm 1) = \frac{1}{2}$	$P(\pm 1) = \frac{1}{2}$
σ_x measurement	$P(+1) = 1$	$P(\pm 1) = \frac{1}{2}$
Purity	1	$\frac{1}{2}$
Origin of probabilities	Superposition	OR rule over B

Same σ_z probabilities; different σ_x probabilities. The difference is the coherences: present in the pure state (amplitudes interfere), absent in the mixed state (the OR sum has erased them). The missing coherence has not disappeared; it has moved into the correlations in the joint state $|\Phi^+\rangle$, which is still pure ($\text{Tr}[\rho_{AB}^2] = 1$).

10.9 Temperature from the tensor product degeneracy

Setup. System A is coupled to a large bath B . The joint state is the ground state $|\Omega\rangle$ of $H = H_A + H_B + H_{\text{int}}$. The observer measures $H_A \otimes \mathbf{1}_B$ with eigenvalues E_a . Each E_a has degeneracy n_B .

The Boltzmann distribution from the OR sum. The projected superposition: $h_t^{(a)} = \sum_b c(a, b) |a, b\rangle$. The Born probability: $P(E_a) = \sum_b |c(a, b)|^2$. For a large bath, the number of bath states compatible with system energy E_a is $W_B(E_{\text{total}} - E_a) = e^{S_B/k_B}$, and the OR sum

gives:

$$P(E_a) \approx \frac{e^{-\beta E_a}}{Z}, \quad \beta = \frac{\partial S_B}{\partial E_B}, \quad \rho_A = \frac{e^{-\beta H_A}}{Z}. \quad (10.20)$$

Temperature is degeneracy. Temperature is the parameter of the OR sum when the degeneracy is astronomical ($n_B \sim e^{10^{23}}$) and the bath multiplicity grows exponentially with energy. A system “at temperature T ” is a system whose energy eigenvalue has multiplicity $W_B = e^{S_B/k_B}$, and the thermal probabilities are the squared lengths of the projected superpositions summed over this enormous degeneracy.

Temperature is not an independent concept from classical thermodynamics grafted onto quantum mechanics. It is a consequence of the tensor product degeneracy rendered through the Born rule.

10.10 Decoherence: the environment provides enormous degeneracy

Decoherence is the degeneracy framework applied at enormous scale.

The system-environment interaction

System S starts in a superposition; environment E in $|e_0\rangle$:

$$h_0 = (\alpha |0\rangle_S + \beta |1\rangle_S) \otimes |e_0\rangle_E. \quad (10.21)$$

The Schrödinger equation evolves the joint state to:

$$h_t = \alpha |0\rangle_S \otimes |e_0(t)\rangle_E + \beta |1\rangle_S \otimes |e_1(t)\rangle_E, \quad (10.22)$$

where $|e_0(t)\rangle$ and $|e_1(t)\rangle$ become increasingly orthogonal as they encode different information about S . The joint state remains a pure superposition: the Schrödinger equation has not collapsed anything.

The projected superpositions

The observer measures $\sigma_z \otimes \mathbf{1}_E$. Eigenvalue $+1$ has degeneracy $\dim \mathcal{H}_E$: every environment state is an eigenvector. The projections:

$$h_t^{(+1)} = \alpha |0\rangle_S \otimes |e_0(t)\rangle_E, \quad h_t^{(-1)} = \beta |1\rangle_S \otimes |e_1(t)\rangle_E. \quad (10.23)$$

Born probabilities: $P(+1) = |\alpha|^2$, $P(-1) = |\beta|^2$ —unchanged by the environment interaction. The diagonal entries of ρ_S are unaffected by decoherence.

The density matrix and the decoherence factor

The reduced density matrix:

$$\rho_S(t) = \begin{pmatrix} |\alpha|^2 & \alpha\beta^*\gamma(t) \\ \alpha^*\beta\gamma^*(t) & |\beta|^2 \end{pmatrix}, \quad \gamma(t) = \langle e_1(t)|e_0(t)\rangle. \quad (10.24)$$

The diagonal entries (Born probabilities) are unaffected. The off-diagonal entries (coherences) are suppressed by $|\gamma(t)|$: the overlap of the environment branches.

Why the coherences decay. The off-diagonal $(\rho_S)_{0,1} = \alpha\beta^*\gamma(t)$ comes from the cross term between the two projected superpositions. In the OR sum over environment states: $\sum_e \langle e|e_0(t)\rangle \langle e_1(t)|e\rangle = \langle e_1(t)|e_0(t)\rangle = \gamma(t)$. As the branches become orthogonal:

$$|\gamma(t)| = e^{-\Gamma t/2} \rightarrow 0, \quad \rho_S \rightarrow \begin{pmatrix} |\alpha|^2 & 0 \\ 0 & |\beta|^2 \end{pmatrix}. \quad (10.25)$$

The density matrix becomes diagonal: a classical probability distribution with no interference. The OR sum over the enormous environmental degeneracy has completed its work.

The pointer basis

The basis in which ρ_S diagonalises is selected by the system-environment interaction. The states most robust to environmental entanglement—those for which $|\gamma(t)|$ decays slowest—are the *pointer states*. For a particle coupled to the environment through position, the pointer basis is approximately the position basis—which is why macroscopic objects have definite locations.

What decoherence achieves and what it does not

Claim	Assessment
Explains classical appearance	Yes: OR sum over $\dim \mathcal{H}_E \sim e^{10^{23}}$ kills coherences
Selects the pointer basis	Yes: interaction determines which basis decoheres
Eliminates superposition	No: superposition moves to joint space
Derives the Born rule	No: Born rule remains a postulate
Derives the Bohr collapse	No: the trim is still needed
Solves the measurement problem	No: Born and Bohr remain irreducible

Decoherence operates entirely within the Schrödinger layer: it is a consequence of unitary evolution. The joint state h_t remains a pure superposition; the superposition has moved to the joint system-environment space, not disappeared. Decoherence explains *why* the Born layer finds classical-looking outcomes (the coherences have decayed), but it does not replace the Born-Bohr rendering layer. The Born-Bohr layer remains the only way to produce classical outcomes from the quantum hidden state.

10.11 The degeneracy hierarchy, unified

Every phenomenon in this section is one instance of the degeneracy framework:

Phenomenon	Observable	Degeneracy source	Consequence
No measurement	$\mathbf{1}_A \otimes \mathbf{1}_B$	Both A and B	Nothing happens
Partial measurement	$\hat{\mathcal{O}}_A \otimes \mathbf{1}_B$	B unobserved ($d = n_B$)	Density matrix, marginals
Entanglement	Same, with c_{ab} entangled	Same	Correlations, Bell violation
Temperature	$H_A \otimes \mathbf{1}_B$, large B	$d \sim e^{10^{23}}$	Boltzmann distribution
Decoherence	$\hat{\mathcal{O}}_S \otimes \mathbf{1}_E$, large E	$d \sim e^{10^{23}}$	Coherences decay
Classical world	Macroscopic collective	$d \sim e^{S_{\text{thermo}}}$	Deterministic output

The tensor product does not introduce new physics. It introduces a new *source* of degeneracy—the unobserved subsystem—that makes the general framework concrete: the identity on B is the maximally degenerate observable from the previous section, applied to the specific context of a composite system. The OR sum over the degenerate index produces classical probabilities; the size of the degeneracy determines whether the output looks quantum (small d) or classical (large d).

The general framework of the previous section asked: what is the degeneracy structure of the observable? The tensor product specialisation asks: which subsystems does the observer not see?

The answer to the second question determines the answer to the first. Every measurement in nature is a restricted measurement—an observable tensored with the identity on everything the observer cannot access—and the degeneracy is always set by the dimension of the unobserved Hilbert space.

11 Angular Momentum: the First Representation

Angular momentum is the simplest and most important example of the algebraic structure that will recur throughout this book: a set of operators satisfying commutation relations that define a Lie algebra, whose representation theory classifies the possible quantum states. It is also the first observable we encounter whose eigenvalues come from algebra alone—not from solving a differential equation.

11.1 The angular momentum algebra

In classical mechanics, the angular momentum of a particle with position $\mathbf{r}$ and momentum $\mathbf{p}$ is $\mathbf{L} = \mathbf{r} \times \mathbf{p}$. In quantum mechanics, $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ are operators satisfying $[\hat{r}_i, \hat{p}_j] = i\hbar\delta_{ij}$. The angular momentum operators are:

$$\hat{L}_i = \epsilon_{ijk}\hat{r}_j\hat{p}_k, \quad i = 1, 2, 3, \quad (11.1)$$

where ϵ_{ijk} is the Levi-Civita symbol. Explicitly: $\hat{L}_1 = \hat{r}_2\hat{p}_3 - \hat{r}_3\hat{p}_2$, and cyclically.

The commutation relations. From $[\hat{r}_i, \hat{p}_j] = i\hbar\delta_{ij}$, a direct computation gives:

$$[\hat{L}_i, \hat{L}_j] = i\hbar\epsilon_{ijk}\hat{L}_k. \quad (11.2)$$

This is the entire content of the angular momentum algebra. The three components $\hat{L}_1$, $\hat{L}_2$, $\hat{L}_3$ do not commute with each other: measuring one disturbs the others. But the total angular momentum squared:

$$\hat{L}^2 = \hat{L}_1^2 + \hat{L}_2^2 + \hat{L}_3^2 \quad (11.3)$$

commutes with every component:

$$[\hat{L}^2, \hat{L}_i] = 0 \quad \text{for all } i. \quad (11.4)$$

Therefore $\hat{L}^2$ and any one component (conventionally $\hat{L}_3$) can be simultaneously diagonalised: they share a common eigenbasis.

The algebra is all we need. The commutation relations (11.2) were derived from $\hat{\mathbf{L}} = \hat{\mathbf{r}} \times \hat{\mathbf{p}}$, but the remarkable fact is that the eigenvalues of $\hat{L}^2$ and $\hat{L}_3$ follow from the commutation relations alone, without any reference to position, momentum, or wavefunctions. This is the power of the algebraic approach: the spectrum is determined by the algebra, not by the representation.

11.2 The eigenvalues: ladder operators

Define the raising and lowering operators:

$$\hat{L}_+ = \hat{L}_1 + i\hat{L}_2, \quad \hat{L}_- = \hat{L}_1 - i\hat{L}_2. \quad (11.5)$$

From (11.2):

$$[\hat{L}_3, \hat{L}_\pm] = \pm\hbar\hat{L}_\pm, \quad [\hat{L}_+, \hat{L}_-] = 2\hbar\hat{L}_3, \quad [\hat{L}^2, \hat{L}_\pm] = 0. \quad (11.6)$$

The first relation is the key: it says that $\hat{L}_+$ *raises* the eigenvalue of $\hat{L}_3$ by $\hbar$, and $\hat{L}_-$ *lowers* it by $\hbar$.

The spectrum. Let $|\ell, m\rangle$ be a simultaneous eigenstate of $\hat{L}^2$ and $\hat{L}_3$:

$$\hat{L}^2 |\ell, m\rangle = \hbar^2\ell(\ell+1) |\ell, m\rangle, \quad \hat{L}_3 |\ell, m\rangle = \hbar m |\ell, m\rangle. \quad (11.7)$$

The ladder operators act as:

$$\hat{L}_+ |\ell, m\rangle = \hbar\sqrt{\ell(\ell+1) - m(m+1)} |\ell, m+1\rangle, \quad \hat{L}_- |\ell, m\rangle = \hbar\sqrt{\ell(\ell+1) - m(m-1)} |\ell, m-1\rangle. \quad (11.8)$$

Derivation from algebra alone. The argument parallels the harmonic oscillator (which we treat in the next section, but the logic is the same).

1. $\hat{L}_3 |\ell, m\rangle = \hbar m |\ell, m\rangle$ and $[\hat{L}_3, \hat{L}_+] = \hbar \hat{L}_+$ together give $\hat{L}_3(\hat{L}_+ |\ell, m\rangle) = \hbar(m+1)(\hat{L}_+ |\ell, m\rangle)$: the state $\hat{L}_+ |\ell, m\rangle$ has $\hat{L}_3$ -eigenvalue $\hbar(m+1)$.
2. Since $[\hat{L}^2, \hat{L}_+] = 0$, the state $\hat{L}_+ |\ell, m\rangle$ has the same $\hat{L}^2$ -eigenvalue $\hbar^2 \ell(\ell+1)$. The ladder moves m but not ℓ .
3. Since $\hat{L}^2 - \hat{L}_3^2 = \hat{L}_1^2 + \hat{L}_2^2 \geq 0$, we have $\ell(\ell+1) \geq m^2$, so $|m| \leq \ell$ (roughly). The ladder must terminate: there exist $m_{\max}$ and $m_{\min}$ such that $\hat{L}_+ |\ell, m_{\max}\rangle = 0$ and $\hat{L}_- |\ell, m_{\min}\rangle = 0$.
4. The termination conditions give $m_{\max} = \ell$ and $m_{\min} = -\ell$. The ladder has $2\ell + 1$ rungs: $m = -\ell, -\ell + 1, \dots, \ell - 1, \ell$.
5. For $2\ell + 1$ to be a positive integer, ℓ must be a non-negative integer or half-integer: $\ell = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots$

The orbital angular momentum (from $\hat{\mathbf{L}} = \hat{\mathbf{r}} \times \hat{\mathbf{p}}$) has integer ℓ only. The half-integer values ($\ell = \frac{1}{2}, \frac{3}{2}, \dots$) correspond to *spin* angular momentum, which has no classical counterpart and cannot be written as $\hat{\mathbf{r}} \times \hat{\mathbf{p}}$. The algebra (11.2) is more general than its orbital realisation: it admits representations that the classical cross product does not.

11.3 The algebra is $\mathfrak{su}(2) \cong \mathfrak{so}(3)$

The commutation relations (11.2) define a Lie algebra. Define the dimensionless generators $J_i = \hat{L}_i / \hbar$:

$$[J_i, J_j] = i\epsilon_{ijk} J_k. \quad (11.9)$$

This is the Lie algebra $\mathfrak{su}(2)$ —the algebra of traceless anti-Hermitian 2×2 matrices (with an i absorbed into the convention). It is also isomorphic to $\mathfrak{so}(3)$, the algebra of infinitesimal rotations in 3D.

The connection to rotations is direct. A rotation by angle ϕ around axis $\hat{n}$ is implemented by the unitary operator:

$$U(\hat{n}, \phi) = e^{-i\phi \hat{n} \cdot \mathbf{J}} = e^{-i\phi(n_1 J_1 + n_2 J_2 + n_3 J_3)}. \quad (11.10)$$

For a rotation around the z -axis by angle ϕ :

$$U(z, \phi) = e^{-i\phi J_3}. \quad (11.11)$$

Acting on an eigenstate: $U(z, \phi) |\ell, m\rangle = e^{-im\phi} |\ell, m\rangle$. Each eigenstate acquires a phase $e^{-im\phi}$ proportional to its m value.

Representations of $\mathfrak{su}(2)$. A representation of the angular momentum algebra is a set of matrices J_1, J_2, J_3 satisfying (25.14), acting on a finite-dimensional vector space. The representation theory (derived above by the ladder argument) gives:

- For each $\ell = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots$, there is exactly one irreducible representation of dimension $2\ell + 1$.
- The basis vectors $|\ell, m\rangle$ for $m = -\ell, \dots, \ell$ are the eigenstates of J_3 in this representation.
- The representation ℓ is called the *spin- ℓ representation*.

ℓ	Dimension	m values	Name	Example
0	1	$\{0\}$	Scalar	Higgs boson, pion
$\frac{1}{2}$	2	$\{-\frac{1}{2}, +\frac{1}{2}\}$	Spinor	Electron, quark, neutrino
1	3	$\{-1, 0, +1\}$	Vector	Photon, $W^\pm$, Z^0
$\frac{3}{2}$	4	$\{-\frac{3}{2}, \dots, +\frac{3}{2}\}$	Spin- $\frac{3}{2}$	Δ baryon, gravitino
2	5	$\{-2, \dots, +2\}$	Tensor	Graviton (hypothetical)

This classification—which follows entirely from the algebra (25.14) and the ladder argument—organises every particle in the Standard Model by its spin. The algebra does not care whether the angular momentum is orbital ($\hat{\mathbf{r}} \times \hat{\mathbf{p}}$, integer ℓ) or intrinsic (spin, half-integer ℓ): the representation theory is the same.

11.4 Spherical harmonics: the position-basis representation

For orbital angular momentum ($\ell \in \mathbb{Z}_{\geq 0}$), the eigenstates $|\ell, m\rangle$ have an explicit representation in the position basis: the *spherical harmonics* $Y_\ell^m(\theta, \varphi)$.

From Dirac kets to functions. The wavefunction is $\psi(\mathbf{r}) = \langle \mathbf{r} | \psi \rangle$. In spherical coordinates $\mathbf{r} = (r, \theta, \varphi)$, the angular part of the eigenstate $|\ell, m\rangle$ is:

$$\langle \theta, \varphi | \ell, m \rangle = Y_\ell^m(\theta, \varphi). \quad (11.12)$$

The spherical harmonic Y_ℓ^m is the position-basis representation of the abstract ket $|\ell, m\rangle$, projected onto the angular coordinates.

Explicit forms. The first few spherical harmonics:

$$Y_0^0 = \frac{1}{\sqrt{4\pi}}, \quad (11.13)$$

$$Y_1^0 = \sqrt{\frac{3}{4\pi}} \cos \theta, \quad Y_1^{\pm 1} = \mp \sqrt{\frac{3}{8\pi}} \sin \theta e^{\pm i\varphi}, \quad (11.14)$$

$$Y_2^0 = \sqrt{\frac{5}{16\pi}} (3 \cos^2 \theta - 1), \quad Y_2^{\pm 1} = \mp \sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{\pm i\varphi}, \quad (11.15)$$

$$Y_2^{\pm 2} = \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{\pm 2i\varphi}. \quad (11.16)$$

The φ -dependence is always $e^{im\varphi}$: this is the eigenvalue equation $\hat{L}_3 Y_\ell^m = \hbar m Y_\ell^m$ in the position basis, since $\hat{L}_3 = -i\hbar \partial / \partial \varphi$. The θ -dependence is a polynomial in $\cos \theta$ and $\sin \theta$ of degree ℓ .

Orthonormality and completeness. The spherical harmonics satisfy:

$$\int_0^{2\pi} d\varphi \int_0^\pi d\theta \sin \theta Y_{\ell'}^{m'*}(\theta, \varphi) Y_\ell^m(\theta, \varphi) = \delta_{\ell'\ell} \delta_{m'm}, \quad (11.17)$$

and:

$$\sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} Y_\ell^{m*}(\theta', \varphi') Y_\ell^m(\theta, \varphi) = \frac{\delta(\theta - \theta') \delta(\varphi - \varphi')}{\sin \theta}. \quad (11.18)$$

These are the position-basis versions of $\langle \ell', m' | \ell, m \rangle = \delta_{\ell'\ell} \delta_{m'm}$ (orthonormality) and $\sum_{\ell, m} |\ell, m\rangle \langle \ell, m| = \mathbf{1}$ (completeness). The spherical harmonics are a complete orthonormal basis for functions on the sphere S^2 .

Spherical harmonics as a Fourier transform on the sphere. Just as the Fourier basis $e^{ikx}/\sqrt{2\pi}$ decomposes functions on the line into plane waves labelled by momentum k , the spherical harmonics decompose functions on the sphere into angular modes labelled by (ℓ, m) . Any function $f(\theta, \varphi)$ on S^2 can be expanded:

$$f(\theta, \varphi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} f_{\ell m} Y_{\ell}^m(\theta, \varphi), \quad f_{\ell m} = \int d\Omega Y_{\ell}^{m*}(\theta, \varphi) f(\theta, \varphi). \quad (11.19)$$

The coefficients $f_{\ell m}$ are the angular momentum components of f : the dense representation of f in the $\{|\ell, m\rangle\}$ basis. The spherical harmonic Y_{ℓ}^m is the one-hot basis function; the coefficient $f_{\ell m}$ is the component.

11.5 Angular momentum as an observable: the Born rule

Angular momentum is an observable: a Hermitian operator with real eigenvalues and orthogonal eigenstates. The Born-Bohr framework applies directly.

Measuring $\hat{L}^2$ and $\hat{L}_3$. Let the hidden state be:

$$|h_t\rangle = \sum_{\ell, m} c_{\ell m}(t) |\ell, m\rangle. \quad (11.20)$$

The projected superposition onto the eigensubspace of $\hat{L}^2$ with eigenvalue $\hbar^2\ell(\ell+1)$ is:

$$h_t^{(\ell)} = \sum_{m=-\ell}^{\ell} c_{\ell m}(t) |\ell, m\rangle. \quad (11.21)$$

This is a superposition of $2\ell+1$ eigenstates—all the m values for a given ℓ . The eigenvalue $\hbar^2\ell(\ell+1)$ has degeneracy $2\ell+1$: the measurement of $\hat{L}^2$ cannot distinguish different m values.

The Born probability is the squared length:

$$P(\ell) = \|h_t^{(\ell)}\|^2 = \sum_{m=-\ell}^{\ell} |c_{\ell m}(t)|^2. \quad (11.22)$$

The sum over m is the classical OR rule: the probability of finding angular momentum quantum number ℓ is the sum of the squared amplitudes over all magnetic quantum numbers m , with no interference between different m values.

The Bohr collapse keeps the projected superposition:

$$h_{t+} = \frac{h_t^{(\ell)}}{\|h_t^{(\ell)}\|} = \frac{\sum_m c_{\ell m} |\ell, m\rangle}{\sqrt{\sum_m |c_{\ell m}|^2}}. \quad (11.23)$$

The total angular momentum is now definite (ℓ); the magnetic quantum number m is still uncertain—the superposition over m survives.

Measuring $\hat{L}_3$ after $\hat{L}^2$. If we subsequently measure $\hat{L}_3$, the projected superposition onto eigenvalue $\hbar m_0$ within the ℓ -subspace is a single term:

$$h_{t++} = \frac{c_{\ell m_0}}{\sqrt{\sum_m |c_{\ell m}|^2}} \cdot \frac{|\ell, m_0\rangle}{|c_{\ell m_0}|/\sqrt{\sum_m |c_{\ell m}|^2}} = |\ell, m_0\rangle. \quad (11.24)$$

No degeneracy remains: the state is fully collapsed to a single eigenvector $|\ell, m_0\rangle$.

Measuring $\hat{L}_3$ alone. If we measure $\hat{L}_3$ without first measuring $\hat{L}^2$, the eigenvalue $\hbar m$ has a different degeneracy: it is shared by all states $|\ell, m\rangle$ with any $\ell \geq |m|$. The projected superposition is:

$$h_t^{(m)} = \sum_{\ell \geq |m|} c_{\ell m}(t) |\ell, m\rangle. \quad (11.25)$$

The OR sum is now over ℓ :

$$P(m) = \sum_{\ell \geq |m|} |c_{\ell m}|^2. \quad (11.26)$$

The collapse keeps the superposition over ℓ : the magnetic quantum number is definite, but the total angular momentum is still uncertain.

Complementarity between $\hat{L}_1$ and $\hat{L}_3$. The operators $\hat{L}_1$ and $\hat{L}_3$ do not commute: $[\hat{L}_1, \hat{L}_3] = -i\hbar\hat{L}_2 \neq 0$. A state that is an eigenstate of $\hat{L}_3$ (one-hot in m) is a superposition (dense) in the eigenbasis of $\hat{L}_1$. Measuring $\hat{L}_3$ precisely makes $\hat{L}_1$ maximally uncertain, and vice versa. This is the angular momentum uncertainty principle—the same complementarity between one-hot and dense representations that we saw for position and momentum.

11.6 What this foreshadows

The angular momentum algebra $[J_i, J_j] = i\epsilon_{ijk}J_k$ is the prototype for every symmetry algebra in this book:

- **Spin** is the $\ell = \frac{1}{2}$ representation: the Pauli matrices $\sigma_i/2$ satisfy $[\sigma_i/2, \sigma_j/2] = i\epsilon_{ijk}\sigma_k/2$. We develop this in the next section on the Pauli equation.
- **The Lorentz group** $SL(2, \mathbb{C})$ has the Lie algebra $\mathfrak{sl}(2, \mathbb{C}) \cong \mathfrak{su}(2) \oplus \mathfrak{su}(2)$: two copies of the angular momentum algebra, labelled (j_L, j_R) . The classification of Lorentz representations is the angular momentum ladder argument applied twice.
- **Isospin** $SU(2)_L$ of the electroweak theory has the same algebra: the weak isospin generators $T^a = \sigma^a/2$ satisfy $[T^a, T^b] = i\epsilon^{abc}T^c$. Left-handed fermions form doublets ($T = \frac{1}{2}$); the Higgs is a doublet; right-handed fermions are singlets ($T = 0$).
- **Colour** $SU(3)_c$ of QCD has a larger algebra $[T^a, T^b] = if^{abc}T^c$ with eight generators, but the representation theory follows the same pattern: irreducible representations classified by highest weight, with ladder operators moving between states.
- **Addition of angular momenta** (Clebsch-Gordan decomposition) is the tensor product of representations: $\ell_1 \otimes \ell_2 = |\ell_1 - \ell_2| \oplus \cdots \oplus (\ell_1 + \ell_2)$. This is the prototype for combining particle quantum numbers in the Standard Model.

The same algebraic structure—generators, commutation relations, ladder operators, irreducible representations—appears at every scale of the theory. The angular momentum algebra is where we first see the pattern. Learning it here means learning it once for the entire book.

11.7 The two-valuedness: orbital angular momentum is not enough

The angular momentum algebra admits half-integer representations ($\ell = \frac{1}{2}, \frac{3}{2}, \dots$), but the orbital angular momentum $\hat{\mathbf{L}} = \hat{\mathbf{r}} \times \hat{\mathbf{p}}$ produces only integer values. The wave equation alone—the Schrödinger equation for $\psi(\mathbf{r}, t)$ —gives orbital quantum numbers $\ell = 0, 1, 2, \dots$. Experiment, however, demands half-integers.

The experimental problem: Stern–Gerlach (1922)

In the Stern–Gerlach experiment, a beam of silver atoms passes through an inhomogeneous magnetic field. If the atoms carry angular momentum ℓ , the beam should split into $2\ell + 1$

components (one for each value of m). For $\ell = 0$: no splitting (one spot). For $\ell = 1$: three spots. For $\ell = 2$: five spots.

The experiment produced *two* spots. Two is $2\ell + 1$ for $\ell = \frac{1}{2}$ —a half-integer value that the orbital angular momentum cannot produce.

The anomalous Zeeman effect (the splitting of spectral lines in a magnetic field) confirmed the problem: the observed splitting patterns required an angular momentum quantum number j that takes half-integer values. No orbital wavefunction $Y_\ell^m(\theta, \varphi)$ can account for this, because the single-valuedness of Y_ℓ^m under $\varphi \rightarrow \varphi + 2\pi$ forces m (and hence ℓ) to be an integer.

The spinning top proposal (1925)

In 1925, George Uhlenbeck and Samuel Goudsmit proposed that the electron possesses an intrinsic angular momentum—as if it were a tiny spinning top rotating about its own axis. This intrinsic angular momentum, which they called *spin*, would have quantum number $s = \frac{1}{2}$, giving $2s + 1 = 2$ orientations and explaining the two spots in Stern–Gerlach.

The word “spin” stuck, but the picture of a spinning top is misleading. A classical spinning sphere of the electron’s size and charge, rotating fast enough to produce the observed magnetic moment, would have surface velocities far exceeding the speed of light. The electron is not spinning in any classical sense. The word “spin” is a historical misnomer that has persisted for a century.

Pauli’s internal index (1927)

Pauli cut through the confusion with a precise mathematical proposal. He did not ask what spin “is” physically—he asked what it requires mathematically.

His answer: the wavefunction of the electron must have *two components*. Instead of a single function $\psi(\mathbf{r}, t)$, the electron is described by a pair:

$$\Psi(\mathbf{r}, t) = \begin{pmatrix} \psi_\uparrow(\mathbf{r}, t) \\ \psi_\downarrow(\mathbf{r}, t) \end{pmatrix}. \quad (11.27)$$

The two components correspond to the two values of the spin projection: $s_z = +\frac{1}{2}\hbar$ (spin up) and $s_z = -\frac{1}{2}\hbar$ (spin down).

In the Dirac notation of this book: the basis vector is not $|\mathbf{r}\rangle$ but $|\mathbf{r}, s\rangle$, where $s \in \{\uparrow, \downarrow\}$ is an internal index. The Hilbert space is the direct product:

$$\mathcal{H} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2 = \mathcal{H}_{\text{space}} \otimes \mathcal{H}_{\text{spin}}. \quad (11.28)$$

This is exactly the direct product we constructed concretely in the Fock section: $|\mathbf{r}\rangle$ lives in the spatial Hilbert space; $|s\rangle$ lives in the two-dimensional spin space $\mathbb{C}^2$; their direct product $|\mathbf{r}, s\rangle$ is a basis vector of the joint space.

The spin operators $\hat{S}_i = \frac{\hbar}{2}\sigma_i$ act on the $\mathbb{C}^2$ factor (the internal index), leaving the spatial part untouched. They satisfy the same algebra $[\hat{S}_i, \hat{S}_j] = i\hbar\epsilon_{ijk}\hat{S}_k$ as the orbital angular momentum—but they act on a different space. The total angular momentum $\hat{\mathbf{J}} = \hat{\mathbf{L}} + \hat{\mathbf{S}}$ combines both.

What Pauli achieved. Pauli did not explain what spin “is” in terms of something more fundamental. He showed that the two-valuedness requires an internal degree of freedom—a two-dimensional index attached to each point in space—and that this index transforms under the $\ell = \frac{1}{2}$ representation of the angular momentum algebra. The Stern–Gerlach two-spot pattern, the anomalous Zeeman effect, and the spin-orbit coupling all follow from the algebra of $\hat{\mathbf{S}}$ acting on $\mathbb{C}^2$.

Spin as rotation in an internal space

Can we recover the legitimacy of the word “spin”? Yes—if we understand what is rotating and where.

A spatial rotation by angle ϕ around the z -axis acts on the orbital part as $e^{-i\phi\hat{L}_3/\hbar}$ and on the spin part as $e^{-i\phi\hat{S}_3/\hbar} = e^{-i\phi\sigma_3/2}$. For the spin part:

$$e^{-i\phi\sigma_3/2} = \begin{pmatrix} e^{-i\phi/2} & 0 \\ 0 & e^{i\phi/2} \end{pmatrix}. \quad (11.29)$$

The spin-up component acquires phase $e^{-i\phi/2}$; the spin-down component acquires phase $e^{+i\phi/2}$. This is a rotation in the *internal two-dimensional space* $\mathbb{C}^2$: the two-component spinor rotates in its internal space, just as a spatial vector rotates in physical space.

So the electron does “spin”—not as a classical top rotating in physical space, but as a quantum object whose internal state rotates in the abstract two-dimensional space $\mathbb{C}^2$ under the action of $SU(2)$. The word “spin” is apt if understood this way: the internal index undergoes a rotation, with the rotation group being $SU(2)$ rather than $SO(3)$.

360° vs. 720°: why spinors are not vectors

Now set $\phi = 2\pi$ (a full 360° rotation) in (11.29):

$$e^{-i\pi\sigma_3} = \begin{pmatrix} e^{-i\pi} & 0 \\ 0 & e^{i\pi} \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbf{1}. \quad (11.30)$$

A 360° rotation sends the spinor $\Psi \rightarrow -\Psi$: it *changes sign*. A spatial wavefunction $\psi(\mathbf{r})$ returns to itself under 360° (because $e^{-i\cdot 2\pi\cdot m} = 1$ for integer m), but a spinor does not. Only a 720° rotation ($\phi = 4\pi$) returns the spinor to itself:

$$e^{-i\cdot 2\pi\sigma_3} = \begin{pmatrix} e^{-2\pi i} & 0 \\ 0 & e^{2\pi i} \end{pmatrix} = +\mathbf{1}. \quad (11.31)$$

Why the sign does not affect single-particle measurements. The Born probability $P(s) = |\Psi_s|^2$ depends on $|\Psi|^2$, not on Ψ itself. The sign change $\Psi \rightarrow -\Psi$ leaves $|\Psi|^2$ unchanged:

$$P(s) = |\Psi_s|^2 = |-\Psi_s|^2. \quad (11.32)$$

A 360° rotation changes the wavefunction but not any measurable property of a single particle. The sign is physically invisible in single-particle measurements—but it becomes visible in interferometry (the neutron 4π experiment) and is essential for the spin-statistics theorem: the (-1) under 360° is what forces identical spin- $\frac{1}{2}$ particles to obey Fermi-Dirac statistics.

The double cover, again. This is the double cover $SU(2) \rightarrow SO(3)$: the two matrices U and $-U$ in $SU(2)$ give the same rotation in $SO(3)$. The spatial rotation group $SO(3)$ has 360° = identity; the internal rotation group $SU(2)$ has 720° = identity. The spinor $\Psi \in \mathbb{C}^2$ transforms under $SU(2)$, not $SO(3)$, and therefore sees the difference between 0° and 360°.

Foreshadowing the Pauli equation

In the next section, we develop the Pauli equation: the Schrödinger equation for the two-component spinor $\Psi(\mathbf{r}, t)$ in the presence of an electromagnetic field. The key features:

- The Hamiltonian acts on $\mathcal{H}_{\text{space}} \otimes \mathbb{C}^2$: it couples the spatial and spin degrees of freedom through the spin-orbit interaction and the Zeeman term $-\hat{\boldsymbol{\mu}} \cdot \mathbf{B}$ (where $\hat{\boldsymbol{\mu}} = -(e/2m_e)\hat{\boldsymbol{\sigma}}$ is the magnetic moment operator).
- The g -factor: the magnetic moment of the electron is not $e/(2m_e)$ (as a classical spinning charge would give) but $g \cdot e/(2m_e)$ with $g \approx 2$ —the Dirac value, later corrected by QED to $g = 2.002\,319\,304\,362\dots$.
- The basis vector $|\mathbf{r}, s\rangle$ is a direct product; the Born rule projects onto subspaces of $\hat{L}^2$, $\hat{S}^2$, $\hat{J}^2$, $\hat{J}_3$ as projected superpositions over the complementary quantum numbers.
- The Pauli equation is the non-relativistic limit of the Dirac equation, which will be treated in the following section.

The internal index s that Pauli introduced—a two-dimensional label attached to each spatial point, transforming under $SU(2)$, changing sign under 360° —will generalise to the spinor indices $m, \bar{m}$ of the Lorentz group, the isospin index of $SU(2)_L$, and the colour index of $SU(3)_c$. Every internal index in the Standard Model is a descendant of Pauli’s spin.

12 Spin and the Pauli Equation

So far the hidden state h_t has been a superposition of position eigenstates $|x\rangle$, with the Schrödinger equation rotating h_t under the Hamiltonian H . A real electron has an additional internal degree of freedom — *spin* — that is not captured by position alone. Incorporating spin requires enriching the basis vectors by adding a spinor index. The Schrödinger layer and Born rendering layer remain unchanged; only the basis vectors and the Hamiltonian become richer.

12.1 Enriched basis vectors: the spinor index

For a spinless particle, the position basis vectors $\{|x\rangle\}$ suffice: the hidden state is $h_t = \int dx \psi(x, t) |x\rangle$, and $\psi(x, t) = \langle x | h_t \rangle$ is the wavefunction.

For a spin- $\frac{1}{2}$ particle such as the electron, each position x must additionally carry a spin label $s \in \{1, 2\}$ (also written $\{\uparrow, \downarrow\}$ or $\{+\frac{1}{2}, -\frac{1}{2}\}$). The basis vectors are now:

$$|x, s\rangle = |x\rangle \otimes |s\rangle, \quad x \in \mathbb{R}^3, \quad s \in \{1, 2\}, \quad (12.1)$$

a direct product of a position eigenstate and a spin basis vector. The orthonormality and completeness relations are:

$$\langle x, s | x', s' \rangle = \delta^3(x - x') \delta_{ss'}, \quad \sum_s \int d^3x |x, s\rangle \langle x, s| = \mathbf{1}. \quad (12.2)$$

The full hidden state is a superposition over all basis vectors $|x, s\rangle$:

$$h_t = \sum_{s=1,2} \int d^3x \psi_s(x, t) |x, s\rangle, \quad \psi_s(x, t) = \langle x, s | h_t \rangle, \quad (12.3)$$

with normalization $\sum_s \int d^3x |\psi_s(x, t)|^2 = 1$. The two functions $\psi_1(x, t)$ and $\psi_2(x, t)$ are the spin-up and spin-down wavefunctions respectively: the probability of finding the electron at x with spin s is $|\psi_s(x, t)|^2 d^3x$.

Equivalently: the spinor as a 2-vector at each position. At each fixed position x , the pair $(\psi_1(x, t), \psi_2(x, t))$ is a two-dimensional complex vector — the *spinor* at x . The hidden state can be visualised as assigning a 2-component spinor to each point in space simultaneously. In the momentum basis $\{|k, s\rangle\}$ (Fourier dual of $\{|x, s\rangle\}$):

$$h_t = \sum_{s=1,2} \int \frac{d^3k}{(2\pi)^3} \tilde{\psi}_s(k, t) |k, s\rangle, \quad \tilde{\psi}_s(k, t) = \langle k, s | h_t \rangle. \quad (12.4)$$

The choice of basis ($\{|x, s\rangle\}$ or $\{|k, s\rangle\}$) is the same basis agnosticism as before; the hidden state h_t is the same object in both representations.

In Dirac notation: the spinor index as a second slot. The wavefunction $\psi_s(x, t) = \langle x, s | h_t \rangle$ has two indices: x (continuous position) and s (discrete spin). The first index is read off by the position basis; the second by the spin basis. The full hidden state is a vector in the tensor product Hilbert space:

$$\mathcal{H} = L^2(\mathbb{R}^3)_x \otimes \mathbb{C}_s^2, \quad (12.5)$$

a continuous position space tensored with a two-dimensional spin space.

12.2 The Hamiltonian acts on both indices

The Schrödinger equation remains:

$$\frac{d}{dt}h_t = -iH h_t, \quad h_t = e^{-iHt}h_0, \quad (12.6)$$

with H now acting on the enriched Hilbert space (12.5). The Hamiltonian can act on the position index x (as before), on the spin index s , or on both simultaneously.

Action on the position index only. The kinetic and potential terms act independently on each spin component, identically for spin-up and spin-down:

$$H_0 |x, s\rangle = \left(-\frac{\nabla^2}{2m} + V(x)\right) |x, s\rangle, \quad (12.7)$$

where the operator $-\nabla^2/2m + V(x)$ acts on the x label and leaves s unchanged. This is a left action (on the position index), blind to spin.

Action on the spin index only. A term that acts on the spin index alone is a 2×2 matrix M in spin space, acting identically at each position x :

$$(M \cdot h_t)_s(x) = \sum_{s'} M_{ss'} \psi_{s'}(x, t). \quad (12.8)$$

The same matrix M is applied to the spinor $(\psi_1(x), \psi_2(x))$ at every x simultaneously.

Mixed action. A term that acts on both indices simultaneously, such as a position-dependent magnetic field coupling to spin, is a product of a position operator and a spin matrix:

$$(B_i(x) \sigma_i \cdot h_t)_s(x) = B_i(x) \sum_{s'} (\sigma_i)_{ss'} \psi_{s'}(x, t). \quad (12.9)$$

The 2×2 spin matrix σ_i and the scalar $B_i(x)$ are multiplied together, coupling both indices.

12.3 Pauli matrices: the spin operators

The operators that act on the spin index $s \in \{1, 2\}$ are 2×2 matrices. The three fundamental spin operators are the *Pauli matrices*:

$$\sigma_1 = \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (12.10)$$

Together with the identity I_2 , they span all 2×2 Hermitian matrices: $M = a_0 I + \vec{a} \cdot \vec{\sigma}$ for any $a_0, \vec{a} \in \mathbb{R}$.

Key properties. Each Pauli matrix is simultaneously:

- **Hermitian:** $\sigma_i^\dagger = \sigma_i$ — a valid spin observable with real eigenvalues ± 1 .
- **Traceless:** $\text{Tr } \sigma_i = 0$ — spin-up and spin-down cancel in the trace.
- **Involutory:** $\sigma_i^2 = I$ — measuring spin twice returns the same result.
- **Unitary:** $\sigma_i^\dagger \sigma_i = I$ — spin rotations preserve the norm.

Commutation and anticommutation. The Pauli matrices satisfy:

$$[\sigma_i, \sigma_j] = 2i \epsilon_{ijk} \sigma_k, \quad (12.11)$$

$$\{\sigma_i, \sigma_j\} = 2\delta_{ij} I, \quad (12.12)$$

where ϵ_{ijk} is the Levi-Civita symbol. Combining:

$$\sigma_i \sigma_j = \delta_{ij} I + i \epsilon_{ijk} \sigma_k. \quad (12.13)$$

The commutation relation (12.11) encodes the non-commutativity of spin measurements along different axes: measuring σ_x then σ_y gives a different result from σ_y then σ_x . The spin operators $S_i = \frac{1}{2} \sigma_i$ satisfy $[S_i, S_j] = i \epsilon_{ijk} S_k$: the angular momentum algebra $\mathfrak{su}(2)$.

Eigenstates. The eigenstates of σ_z are the spin-up and spin-down states:

$$\sigma_z |\uparrow\rangle = +|\uparrow\rangle, \quad |\uparrow\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \sigma_z |\downarrow\rangle = -|\downarrow\rangle, \quad |\downarrow\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \quad (12.14)$$

The eigenstates of σ_x are $|\pm x\rangle = \frac{1}{\sqrt{2}}(1, \pm 1)^T$ with eigenvalues ± 1 ; those of σ_y are $|\pm y\rangle = \frac{1}{\sqrt{2}}(1, \pm i)^T$.

Rotations from the exponential. Since $(\hat{n} \cdot \vec{\sigma})^2 = I$ for any unit vector $\hat{n}$ (from (12.12)), the matrix exponential collapses to:

$$e^{-i\frac{\theta}{2}\hat{n}\cdot\vec{\sigma}} = I \cos \frac{\theta}{2} - i(\hat{n} \cdot \vec{\sigma}) \sin \frac{\theta}{2}. \quad (12.15)$$

This is an $SU(2)$ rotation: it rotates the spinor by angle θ around axis $\hat{n}$. The factor of $\frac{1}{2}$ means a full 2π rotation gives $e^{-i\pi\hat{n}\cdot\vec{\sigma}} = -I$: the spinor acquires a minus sign under a full rotation. A 4π rotation is required to return the spinor to itself. This sign is physically observable in neutron interferometry and is the origin of Fermi-Dirac statistics.

12.4 The Pauli Hamiltonian motor

For an electron (charge $-e$, mass m) in an electromagnetic field with vector potential $\vec{A}(x)$, scalar potential $\phi(x)$, and magnetic field $\vec{B} = \nabla \times \vec{A}$, the *Pauli Hamiltonian* is:

$$H_{\text{Pauli}} = \frac{1}{2m}(\vec{P} + e\vec{A}(X))^2 - \frac{e}{2m}\vec{B}(X) \cdot \vec{\sigma} - e\phi(X). \quad (12.16)$$

Acting on the spinor wavefunction $\Psi(x, t) = (\psi_1(x, t), \psi_2(x, t))^T$:

$$i\partial_t \Psi = \left[\frac{(\vec{P} + e\vec{A})^2}{2m} - e\phi \right] \Psi - \frac{e}{2m} \vec{B} \cdot \vec{\sigma} \Psi. \quad (12.17)$$

The first term acts on position only (same on both spin components); the second term $\vec{B} \cdot \vec{\sigma}$ is a 2×2 spin matrix that mixes the two components.

Physical consequences.

Larmor precession. In a uniform magnetic field $\vec{B} = B\hat{z}$ with no other interactions, the Hamiltonian reduces to a pure spin term:

$$H = -\frac{eB}{2m}\sigma_z = \frac{\omega_L}{2}\sigma_z, \quad \omega_L = -\frac{eB}{m}. \quad (12.18)$$

The Schrödinger evolution gives:

$$\Psi(x, t) = e^{-i\omega_L \sigma_z t/2} \Psi(x, 0) = \begin{pmatrix} e^{-i\omega_L t/2} & 0 \\ 0 & e^{+i\omega_L t/2} \end{pmatrix} \begin{pmatrix} \psi_1(x, 0) \\ \psi_2(x, 0) \end{pmatrix}. \quad (12.19)$$

The spin-up and spin-down components acquire opposite phases, causing the spinor to precess around $\hat{z}$ at the *Larmor frequency* ω_L . This is a rotation of the spin index at every point x , with the position wavefunction completely unaffected.

Zeeman splitting. The energy levels split in a magnetic field. For an atom with orbital wavefunction $\phi(x)$, the two spin states have energies:

$$E_{\pm} = E_0 \pm \frac{eB}{2m}, \quad (12.20)$$

where E_0 is the zero-field energy. The energy gap $\Delta E = eB/m$ is proportional to the field strength. The discrete rotation rates in the Schrödinger RNN become spin-dependent: the two energy eigenstates rotate at different frequencies, and their interference produces the Larmor precession.

Stern-Gerlach experiment. A particle with spin state $|\chi\rangle = \cos(\theta/2)|\uparrow\rangle + \sin(\theta/2)|\downarrow\rangle$ measured by σ_z gives:

$$p(+\tfrac{1}{2}) = \cos^2 \frac{\theta}{2}, \quad p(-\tfrac{1}{2}) = \sin^2 \frac{\theta}{2}. \quad (12.21)$$

The Born rendering layer projects the spin index onto the σ_z eigenbasis; the position wavefunction factors out completely. After measurement, the Bohr layer resets the spin index to $|\uparrow\rangle$ or $|\downarrow\rangle$ while leaving the position wavefunction unchanged.

The g -factor. The Pauli Hamiltonian predicts a magnetic moment of exactly one Bohr magneton $\mu_B = e/2m$ for the electron — equivalently, a g -factor of $g = 2$. This is a non-trivial prediction: classically, orbital motion gives $g = 1$. The Pauli equation gives $g = 2$ by fiat (the factor of 2 in front of the spin magnetic moment term), which matches experiment to leading order. The small correction $g - 2 \approx 0.00232$ is a quantum field theory effect computed from Feynman diagrams in QED, involving the interaction of the electron with its own radiation field.

12.5 Direct product operators acting on direct product kets

The basis vector $|\mathbf{x}, s\rangle = |\mathbf{x}\rangle \otimes |s\rangle$ lives in the tensor product Hilbert space $\mathcal{H} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$. Operators on $\mathcal{H}$ can be of three kinds: they can act purely on the position index $\mathbf{x}$, purely on the spinor index s , or on both. The cleanest way to handle all three cases is to write operators as tensor products and apply the tensor product rule for their action on tensor product states.

The tensor product rule. If A is an operator on $L^2(\mathbb{R}^3)$ and B is an operator on $\mathbb{C}^2$, their tensor product $A \otimes B$ is an operator on $\mathcal{H} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$, acting on a product ket by:

$$(A \otimes B)(|\mathbf{x}\rangle \otimes |s\rangle) = (A|\mathbf{x}\rangle) \otimes (B|s\rangle). \quad (12.22)$$

Each factor acts on its own space independently. By linearity, the action extends to any superposition:

$$(A \otimes B) \sum_s \int d^3\mathbf{x} \psi_s(\mathbf{x}) |\mathbf{x}, s\rangle = \sum_{s,s'} \int d^3\mathbf{x} (A\psi_{s'})(\mathbf{x}) B_{ss'} |\mathbf{x}, s\rangle, \quad (12.23)$$

where $B_{ss'} = \langle s|B|s'\rangle$ are the matrix elements of B in the spin basis, and $A\psi_{s'}$ denotes the action of A on the function $\psi_{s'}(\mathbf{x})$.

The three types of operator, written as tensor products. Every operator on $\mathcal{H}$ can be written in terms of tensor products:

Position-only operators act on the $L^2(\mathbb{R}^3)$ factor and leave the spinor unchanged. They take the form $A \otimes I_2$, where I_2 is the 2×2 identity on $\mathbb{C}^2$:

$$(A \otimes I_2) |\mathbf{x}, s\rangle = (A |\mathbf{x}\rangle) \otimes (I_2 |s\rangle) = (A |\mathbf{x}\rangle) \otimes |s\rangle. \quad (12.24)$$

Examples: kinetic energy $\frac{\hat{\mathbf{p}}^2}{2m} \otimes I_2$, potential $V(\hat{\mathbf{x}}) \otimes I_2$, orbital angular momentum $\hat{L}_i \otimes I_2$. The $\otimes I_2$ is often suppressed in notation, but it is always implicitly present to match the dimension of the full Hilbert space.

Spin-only operators act on the $\mathbb{C}^2$ factor and leave the position wavefunction unchanged. They take the form $I_\infty \otimes B$, where I_∞ is the identity on $L^2(\mathbb{R}^3)$:

$$(I_\infty \otimes B) |\mathbf{x}, s\rangle = |\mathbf{x}\rangle \otimes (B |s\rangle) = \sum_{s'} B_{ss'} |\mathbf{x}, s'\rangle. \quad (12.25)$$

Examples: $I_\infty \otimes \sigma_i$ (spin measurement along axis i), $I_\infty \otimes e^{-i\theta \hat{n} \cdot \vec{\sigma}/2}$ (spin rotation).

Mixed operators couple both indices and cannot be written as a single tensor product, but as a *sum* of tensor products:

$$\hat{\mathbf{B}}(\hat{\mathbf{x}}) \cdot \vec{\sigma} = \sum_{i=1}^3 B_i(\hat{\mathbf{x}}) \otimes \sigma_i, \quad (12.26)$$

where $B_i(\hat{\mathbf{x}})$ acts on the position factor and σ_i acts on the spin factor. At each point $\mathbf{x}$, the magnetic field $\mathbf{B}(\mathbf{x})$ is a different vector, so the spin rotation it generates depends on position.

The Pauli Hamiltonian as a sum of tensor products. Written explicitly in tensor product form, the Pauli Hamiltonian (12.16) is:

$$H_{\text{Pauli}} = \underbrace{\frac{(\hat{\mathbf{p}} + e\hat{\mathbf{A}})^2}{2m} \otimes I_2 - e\hat{\phi} \otimes I_2}_{\text{position-only: kinetic and scalar potential}} - \underbrace{\frac{e}{2m} \sum_{i=1}^3 B_i(\hat{\mathbf{x}}) \otimes \sigma_i}_{\text{mixed: magnetic coupling}}. \quad (12.27)$$

The $\otimes I_2$ factors in the kinetic and potential terms are usually suppressed in the conventional notation $\hat{\mathbf{p}}^2/2m - e\phi$, but they are essential for dimensional consistency: without them, these operators would act on $L^2(\mathbb{R}^3)$ alone rather than on the full space $L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$. The explicit tensor product form makes clear that the kinetic energy operator is completely blind to the spinor index — it does the same thing to the spin-up component and the spin-down component. The magnetic coupling $\sum_i B_i(\hat{\mathbf{x}}) \otimes \sigma_i$ is the only term that mixes the two spinor components.

Action on the wavefunction. Acting on the spinor wavefunction $\Psi(\mathbf{x}) = \begin{pmatrix} \psi_1(\mathbf{x}) \\ \psi_2(\mathbf{x}) \end{pmatrix}$:

$$H_{\text{Pauli}} \Psi(\mathbf{x}) = \begin{pmatrix} \left[\frac{(\hat{\mathbf{p}} + e\hat{\mathbf{A}})^2}{2m} - e\phi \right] \psi_1(\mathbf{x}) \\ \left[\frac{(\hat{\mathbf{p}} + e\hat{\mathbf{A}})^2}{2m} - e\phi \right] \psi_2(\mathbf{x}) \end{pmatrix} - \frac{e}{2m} \begin{pmatrix} B_z & B_x - iB_y \\ B_x + iB_y & -B_z \end{pmatrix} \begin{pmatrix} \psi_1(\mathbf{x}) \\ \psi_2(\mathbf{x}) \end{pmatrix}, \quad (12.28)$$

where the first term acts identically on each spinor component (the position operator acts on $\mathbf{x}$, the identity acts on the spinor index), and the second term mixes the two components through the matrix $\mathbf{B} \cdot \vec{\sigma}$.

12.6 Total angular momentum: spin is not a spinning top

The orbital angular momentum $\hat{\mathbf{L}} = \hat{\mathbf{x}} \times \hat{\mathbf{p}}$ acts on the position factor $L^2(\mathbb{R}^3)$ and has integer eigenvalues $m_\ell \in \{-\ell, \dots, +\ell\}$ for $\ell = 0, 1, 2, \dots$. The spin angular momentum $\hat{\mathbf{S}} = \frac{1}{2}\vec{\sigma}$ acts on the spin factor $\mathbb{C}^2$ and has half-integer eigenvalues $m_s \in \{-\frac{1}{2}, +\frac{1}{2}\}$.

These are operators on *different* Hilbert spaces, and to add them we must first embed each in the full tensor product space $\mathcal{H} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$ using the tensor product with the identity:

$$\hat{L}_i \text{ on } \mathcal{H} = \hat{L}_i \otimes I_2, \quad (12.29)$$

$$\hat{S}_i \text{ on } \mathcal{H} = I_\infty \otimes \frac{\sigma_i}{2}. \quad (12.30)$$

The total angular momentum is their sum:

$$\hat{J}_i = \hat{L}_i \otimes I_2 + I_\infty \otimes \frac{\sigma_i}{2}. \quad (12.31)$$

The two terms commute trivially: $[\hat{L}_i \otimes I_2, I_\infty \otimes \sigma_j/2] = 0$, because they act on different tensor factors. The total angular momentum satisfies the same algebra as each of its parts:

$$[\hat{J}_i, \hat{J}_j] = i\varepsilon_{ijk} \hat{J}_k. \quad (12.32)$$

The identity factors are essential, not cosmetic. The embedding (12.29)–(12.30) is not a notational convenience. Without the $\otimes I_2$ and $I_\infty \otimes$, the two terms in (12.31) would act on different Hilbert spaces and their sum would be undefined. The identity factors match the dimensions of the two spaces so that $\hat{J}_i$ is a well-defined operator on the same Hilbert space $\mathcal{H} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$ as the Hamiltonian.

Spin is not a spinning top. The spin angular momentum $I_\infty \otimes \frac{1}{2}\sigma_i$ is an intrinsic property of the particle, not a description of any physical rotation of matter about an axis. The electron has no known spatial extent at the scale where spin becomes relevant, and the classical image of a spinning charged sphere — which Pauli and others considered and rejected early in the development of quantum mechanics — leads to inconsistencies: the surface of the electron would need to move faster than light to generate the observed magnetic moment. Spin is instead a *representation-theoretic* property: it labels the transformation of the spinor index $s \in \{1, 2\}$ under rotations of the coordinate system. Under a spatial rotation by angle θ about axis $\hat{\mathbf{n}}$, the orbital factor transforms by $e^{-i\theta\hat{L}_{\hat{\mathbf{n}}}}$ and the spin factor transforms by $e^{-i\theta\sigma_{\hat{\mathbf{n}}}/2}$. The half-integer eigenvalues of $\hat{S}_i$ and the 4π periodicity of the spinor are consequences of the representation theory of the rotation group $SO(3)$, not of any classical rotational motion.

In the Fock-Born language: spin is an index on the monosemantic basis vectors $|\mathbf{x}, s\rangle$, acted on by the 2×2 operator $I_\infty \otimes \sigma_i/2$. The Born rendering of the spin observable $I_\infty \otimes \sigma_z/2$ on the spinor state $\Psi(\mathbf{x})$ gives outcomes $\pm\frac{1}{2}$ with Born probabilities $\int |\psi_\pm(\mathbf{x})|^2 d^3\mathbf{x}$. The spin exists as an index of the hidden state h_t ; it is rendered as a discrete classical label $\pm\frac{1}{2}$ by the Born layer. There is nothing spinning.

Eigenvalues of total angular momentum. The eigenstates of $\hat{J}^2$ and $\hat{J}_z$ are labelled by the total angular momentum quantum number j and the magnetic quantum number $m_j \in \{-j, \dots, +j\}$. For a system with orbital angular momentum ℓ and spin $s = \frac{1}{2}$, the Clebsch-Gordan decomposition gives:

$$\mathcal{H}^{(\ell)} \otimes \mathbb{C}^2 = \mathcal{H}^{(j=\ell+\frac{1}{2})} \oplus \mathcal{H}^{(j=\ell-\frac{1}{2})}, \quad (12.33)$$

a direct sum of two total angular momentum sectors (for $\ell \geq 1$; for $\ell = 0$ only $j = \frac{1}{2}$ survives). The half-integer value $j = \ell \pm \frac{1}{2}$ is the direct consequence of combining integer orbital angular momentum ℓ (from the position factor) with half-integer spin $\frac{1}{2}$ (from the spinor factor) via the addition formula (12.31). The half-integer j cannot arise from any operator acting only on the position factor $L^2(\mathbb{R}^3)$.

12.7 The hidden state as an $\infty \times 2$ matrix: two-sided operator action

There is a particularly clean way to visualise the structure of the spinor hidden state and the action of tensor product operators: the hidden state can be represented as an $\infty \times 2$ matrix, with rows indexed by position $\mathbf{x}$ and columns indexed by spin $s \in \{1, 2\}$:

$$\Psi = \begin{pmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) \\ \vdots & \vdots \end{pmatrix} = (\psi_1(\mathbf{x}) \mid \psi_2(\mathbf{x})), \quad (12.34)$$

where the $\mathbf{x}$ index is continuous (the matrix has infinitely many rows) and the spinor index $s \in \{1, 2\}$ gives two columns. The entry $\Psi_{\mathbf{x}s} = \psi_s(\mathbf{x})$ is the amplitude to find the particle at position $\mathbf{x}$ with spin s . The normalization condition:

$$\|\Psi\|_F^2 = \sum_s \int d^3\mathbf{x} |\psi_s(\mathbf{x})|^2 = 1 \quad (12.35)$$

is the Frobenius norm of the matrix Ψ .

Position-only operators: left multiplication. An operator A acting only on the position index corresponds to *left multiplication* of the matrix Ψ by A :

$$(A \otimes I_2) \Psi = A \Psi, \quad (12.36)$$

where A is the (infinite-dimensional) matrix representation of the operator in the position basis, and $A\Psi$ is matrix multiplication: $(A\Psi)_{\mathbf{x}s} = \sum_{\mathbf{x}'} A_{\mathbf{x}\mathbf{x}'} \Psi_{\mathbf{x}'s}$ (or, in the continuous case, $\int d^3\mathbf{x}' A(\mathbf{x}, \mathbf{x}') \psi_s(\mathbf{x}')$). The column index s is carried along unchanged: A acts identically on each column of Ψ .

Spin-only operators: right multiplication by the transpose. An operator B acting only on the spin index corresponds to *right multiplication* of Ψ by B^T (the transpose):

$$(I_\infty \otimes B) \Psi = \Psi B^T. \quad (12.37)$$

The transpose arises from the index convention. In the matrix Ψ , the spin index s labels *columns*. The operator B acts on the spin ket $|s\rangle$, which is a column vector in $\mathbb{C}^2$; but in the matrix Ψ , the spin index runs along the *row* direction of the product. More precisely:

$$[(I_\infty \otimes B)\Psi]_{\mathbf{x}s} = \sum_{s'} \Psi_{\mathbf{x}s'} B_{s's} = \sum_{s'} \Psi_{\mathbf{x}s'} (B^T)_{ss'} = [\Psi B^T]_{\mathbf{x}s}. \quad (12.38)$$

For Hermitian operators $B = B^\dagger$, the transpose is $B^T = B^*$ (complex conjugate). For real symmetric matrices (such as σ_x and σ_z), $B^T = B$, so right multiplication by B is the same as right multiplication by B^T . For $\sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$, the transpose is $\sigma_y^T = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} = -\sigma_y$, so the right-multiplication form acquires a sign.

Mixed operators: left and right simultaneously. A mixed operator $A \otimes B$ acts by left multiplication by A and right multiplication by B^T :

$$(A \otimes B) \Psi = A \Psi B^T. \quad (12.39)$$

This is the *two-sided matrix multiplication* representation of tensor product operators: the position operator acts on the row index from the left, and the spin operator acts on the column index from the right. The two actions commute because matrix multiplication associates: $A(\Psi B^T) = (A\Psi) B^T$. This is why position and spin operators always commute: they act on different indices of the same matrix, one from the left and one from the right.

Examples in the matrix picture.

- **Kinetic energy** $((-\nabla^2/2m) \otimes I_2)\Psi = (-\nabla^2/2m)\Psi$: the Laplacian acts on each column of Ψ independently (left multiplication by $-\nabla^2/2m$, right multiplication by $I_2^T = I_2$).
- **Spin measurement** $(I_\infty \otimes \sigma_z)\Psi = \Psi \sigma_z^T = \Psi \sigma_z$: since $\sigma_z^T = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, this right-multiplies the matrix by σ_z , flipping the sign of the second column: $(\psi_1, \psi_2) \rightarrow (\psi_1, -\psi_2)$.
- **Larmor precession** $(I_\infty \otimes e^{-i\omega_L \sigma_z t/2})\Psi = \Psi e^{-i\omega_L \sigma_z^T t/2} = \Psi \begin{pmatrix} e^{-i\omega_L t/2} & 0 \\ 0 & e^{+i\omega_L t/2} \end{pmatrix}$: right-multiplies each row (each spinor $\psi(\mathbf{x})$) by the diagonal rotation matrix, causing the two columns to rotate in opposite directions — the spinor precesses at the Larmor frequency at every position $\mathbf{x}$ simultaneously.
- **Magnetic coupling** $(B_z(\hat{\mathbf{x}}) \otimes \sigma_z)\Psi = B_z(\hat{\mathbf{x}})\Psi \sigma_z^T$: left-multiplies by the position-dependent scalar $B_z(\mathbf{x})$ (scaling each row by the local field strength) and right-multiplies by σ_z (flipping the sign of the second column).

Total angular momentum in the matrix picture. The total angular momentum (12.31) acts as:

$$\hat{J}_i \Psi = \hat{L}_i \Psi + \Psi \frac{\sigma_i^T}{2}, \quad (12.40)$$

a two-sided action: the orbital part $\hat{L}_i$ acts on the row index from the left (rotating the position wavefunction), and the spin part $\sigma_i^T/2$ acts on the column index from the right (rotating the spinor). For $\sigma_x^T = \sigma_x$ and $\sigma_z^T = \sigma_z$ (real symmetric), the transpose is transparent. For $\sigma_y^T = -\sigma_y$, the right-multiplication gives a sign: $\hat{J}_y \Psi = \hat{L}_y \Psi - \Psi \sigma_y/2$. The generator of a full spatial rotation by angle θ about axis $\hat{\mathbf{n}}$ is therefore:

$$\Psi \mapsto e^{-i\theta \hat{L}_{\hat{\mathbf{n}}}} \Psi e^{-i\theta(\hat{\mathbf{n}} \cdot \vec{\sigma})^T/2} = e^{-i\theta \hat{L}_{\hat{\mathbf{n}}}} \Psi e^{+i\theta \hat{\mathbf{n}} \cdot \vec{\sigma}/2} \quad (12.41)$$

for $\hat{\mathbf{n}} = \hat{z}$ or $\hat{x}$ (where $\sigma_z^T = \sigma_z$ and $\sigma_x^T = \sigma_x$), and with a sign adjustment for $\hat{\mathbf{n}} = \hat{y}$. The two exponentials act on opposite indices of the matrix Ψ , confirming that orbital and spin rotations are completely independent and commute.

Connection to the spinor section. The two-sided multiplication $A\Psi B^T$ is the general form of a tensor product operator acting on this matrix: left factors act on the infinite row index (position), right factors act on the two-column index (spin). This structure extends naturally to the Dirac spinor ($\infty \times 4$ matrix), where the 4×4 gamma matrices act on the column index from the right, and to the Standard Model fields, where the additional internal indices (colour, isospin) add further columns acted on by the corresponding gauge group matrices.

13 The Dirac Equation: Spin from Spacetime

13.1 The Dirac equation: why nature keeps it simple

Pauli's equation added spin to quantum mechanics by hand: attach a two-component spinor index $s \in \{1, 2\}$ to the wavefunction, couple it to the magnetic field via the Pauli matrices σ_i , and adjust the gyromagnetic ratio g to match experiment ($g = 2$, to leading order). The equation works. But Dirac found it deeply unsatisfying.

Dirac's objection was not technical but aesthetic—and, as it turned out, prophetic. Why would nature introduce spin as an additional structure, bolted onto the Schrödinger equation from outside? If quantum mechanics is as economical as Dirac believed it should be, spin must not be an extra input. It must *emerge* from a deeper structural requirement—a requirement so minimal that it leaves no room for alternatives.

Dirac identified the requirement: the wave equation must be compatible with special relativity. The Schrödinger equation is first order in time ($i\partial_t\psi = H\psi$) but second order in space (through $H = -\nabla^2/2m + V$). This asymmetry between time and space is tolerable in non-relativistic physics, where time and space play different roles. But special relativity puts time and space on equal footing: a Lorentz transformation mixes t and $\mathbf{x}$, and any equation that treats them asymmetrically cannot be Lorentz invariant. The Schrödinger equation is not Lorentz invariant. Neither is the Pauli equation: it adds spin, but the time-space asymmetry remains.

Schrödinger himself had encountered this problem before publishing his equation. In late 1925, he wrote down a relativistic wave equation—what we now call the Klein–Gordon equation:

$$\left(\frac{\partial^2}{\partial t^2} - \nabla^2 + m^2\right)\psi = 0, \quad (13.1)$$

which is second order in both time and space, Lorentz invariant, and consistent with the relativistic energy-momentum relation $E^2 = \mathbf{p}^2 + m^2$. But Schrödinger found that it gave the wrong fine structure of the hydrogen spectrum (because it does not include spin), and he discarded it in favour of the non-relativistic first-order equation that bears his name. Klein and Gordon later published the same equation independently.

At a meeting in the late 1920s, Bohr asked Dirac what he was working on. Dirac replied that he was trying to find a relativistic wave equation—a Schrödinger equation that is Lorentz invariant. Bohr pointed out that the Klein–Gordon equation already achieved this. But Dirac was unsatisfied: the Klein–Gordon equation is second order in time. The Schrödinger equation is first order in time, and Dirac believed this was essential—not a historical accident but a structural requirement of quantum mechanics (the first-order time derivative is what makes $|\psi|^2$ a conserved probability density and the evolution unitary).

Since Lorentz invariance puts time and space on equal footing, and the equation must be first order in time, Dirac concluded: it must also be first order in space. He sought an equation of the form:

$$i\frac{\partial\psi}{\partial t} = (-i\boldsymbol{\alpha} \cdot \nabla + \beta m)\psi, \quad (13.2)$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and β are objects to be determined by the requirement that the equation be consistent with $E^2 = \mathbf{p}^2 + m^2$. Squaring both sides and demanding $E^2 = \mathbf{p}^2 + m^2$ forces:

$$\{\alpha_i, \alpha_j\} = 2\delta_{ij}, \quad \{\alpha_i, \beta\} = 0, \quad \beta^2 = 1. \quad (13.3)$$

These are anticommutation relations. No ordinary numbers satisfy them. The smallest matrices that do are 4×4 .

This is Dirac's great discovery: to write a first-order, Lorentz invariant wave equation, the wavefunction must have *four components*. The Schrödinger equation has one component; the

Pauli equation has two; the Dirac equation has four. The two extra components are not put in by hand: they are forced by the requirement of Lorentz invariance combined with first-order time evolution. Spin is not an input; it is an output.

The consequences of this one structural requirement are extraordinary:

- **Spin $\frac{1}{2}$ emerges.** The four-component Dirac spinor decomposes into two two-component Pauli spinors (“large” and “small” components). In the non-relativistic limit, the Dirac equation reduces to the Pauli equation—and the gyromagnetic ratio $g = 2$ emerges automatically, without being postulated. Pauli had to put $g = 2$ in by hand; Dirac gets it for free.
- **Antimatter emerges.** The four-component spinor has solutions with both positive and negative energy: $E = \pm\sqrt{\mathbf{p}^2 + m^2}$. The negative-energy solutions cannot be discarded (they are needed for completeness of the Hilbert space). Dirac initially interpreted them through the “hole theory”: the vacuum is a filled sea of negative-energy states (the Dirac sea), and a hole in this sea is an antiparticle—a positron, with the same mass as the electron but opposite charge. Anderson discovered the positron experimentally in 1932, four years after Dirac’s prediction.
- **A new representation of the Lorentz group emerges.** The four-component object ψ does not transform as a vector under Lorentz transformations (vectors have four components too, but they transform differently). It transforms as a *spinor*: a new, previously unknown representation of the Lorentz group, the $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation. Spinors rotate by half-angles: a 360° rotation gives $\psi \rightarrow -\psi$, and only a 720° rotation returns $\psi \rightarrow \psi$. This representation was not known to mathematicians before Dirac discovered it physically.

The negative-energy problem—the most troubling consequence of the Dirac equation—was eventually resolved not by the hole theory but by the reconceptualisation that Pauli championed in the early 1930s: the Dirac equation is not the quantum mechanics of a single relativistic electron. It is a *classical field equation*, analogous to Maxwell’s equations for the electromagnetic field. Just as Maxwell’s equations are the classical equations of motion for the photon field $A_\mu(x)$, the Dirac equation is the classical equation of motion for the electron field $\psi_A(x)$. Quantization (building the Fock space, promoting ψ to a creation/annihilation operator) comes afterward, as a separate step.

In this reconceptualisation, the negative-energy solutions are not negative-energy electrons: they are the creation operators for positrons. The Dirac sea is replaced by the Fock vacuum; the holes are replaced by antiparticle creation operators; and the equation $i\partial_t\psi = (-i\boldsymbol{\alpha} \cdot \nabla + \beta m)\psi$ is not the Schrödinger equation for a particle but the field equation for the electron field, whose quantization produces the Fock space of electrons and positrons. The same reconceptualisation applies to the Klein–Gordon equation: it is not the quantum mechanics of a relativistic spinless particle but the classical field equation for a scalar field, quantized into a Fock space of scalar particles and antiparticles.

What Dirac discovered, in the end, was not merely an equation. He discovered a new equivariant representation of the Lorentz group—the spinor—that nature uses for all matter. Every quark, every lepton, every fermion in the Standard Model is a Dirac spinor (or, more precisely, a pair of Weyl spinors that combine into a Dirac spinor when mass is present). The requirement that produced this discovery was minimal: a first-order wave equation compatible with Lorentz invariance. The consequences—spin, antimatter, the spinor representation—were maximal. This is Dirac’s beauty: the simplest source code producing the richest emergence.

The following subsections develop the equation, the spinor, the non-relativistic limit, the negative-energy solutions, and the Pauli reconceptualisation in technical detail.

13.2 Special relativity: the coupling of time and space

13.2.1 Where special relativity comes from

Special relativity was forced on physics by electromagnetism. Maxwell's equations in vacuum predict electromagnetic waves traveling at speed $c = 1/\sqrt{\mu_0\epsilon_0}$. But the speed of what, relative to what frame? Newtonian mechanics has a distinguished class of inertial frames related by Galilean transformations:

$$t' = t, \quad \mathbf{x}' = \mathbf{x} - \mathbf{v}t. \quad (13.4)$$

Under Galilean transformation, velocities add: if a wave has speed c in one frame, it has speed $c - v$ in a frame moving at velocity v . But this contradicts experiment — the Michelson-Morley experiment (1887) found no variation in the speed of light regardless of the Earth's motion.

The resolution: Galilean transformations are wrong. The correct transformation law that leaves the speed of light invariant is the **Lorentz transformation**. It mixes time and space:

$$t' = \gamma(t - vx/c^2), \quad (13.5)$$

$$x' = \gamma(x - vt), \quad (13.6)$$

$$y' = y, \quad z' = z, \quad (13.7)$$

where $\gamma = 1/\sqrt{1 - v^2/c^2}$ is the Lorentz factor. The key feature: time and position are *mixed* by the transformation. They are not separate invariant quantities; only their combination is invariant. This is the origin of the 4-dimensional spacetime of special relativity.

The invariant interval. What is preserved under Lorentz transformations is not t alone or $\mathbf{x}$ alone, but the **spacetime interval**:

$$ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2 = \eta_{\mu\nu} dx^\mu dx^\nu, \quad (13.8)$$

where $\eta_{\mu\nu} = \text{diag}(-1, +1, +1, +1)$ is the Minkowski metric. A Lorentz transformation is precisely any linear map on (t, x, y, z) that preserves this quadratic form. These form the Lorentz group $O(1, 3)$.

Why EM is Lorentz invariant. Maxwell's equations were already Lorentz invariant before Einstein; Einstein merely recognized what this implied. The electromagnetic field tensor

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \quad F = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{pmatrix}, \quad (13.9)$$

is an antisymmetric 2-tensor: it transforms under Lorentz transformations as $F_{\mu\nu} \rightarrow \Lambda^\alpha_\mu \Lambda^\beta_\nu F_{\alpha\beta}$. Maxwell's equations $\partial_\mu F^{\mu\nu} = J^\nu$ are Lorentz-covariant equations between tensors, hence automatically invariant. Electric and magnetic fields mix under boosts: what looks like a pure electric field in one frame is a combination of electric and magnetic in a boosted frame.

13.2.2 Tensors: objects that know about coordinate change

The word “tensor” appears throughout this book and throughout physics. Before seeing how tensors marry time with space and energy with momentum, we need to understand what a tensor is and why it is the natural language for physical laws.

The core idea: equivariance under coordinate change. A physical law must be independent of which coordinate system we use to express it. Temperature is 300 K whether we measure it in Celsius or Fahrenheit; the force on a particle is the same vector regardless of whether we use Cartesian or polar coordinates. A tensor is a mathematical object that transforms in a *controlled, predictable* way under coordinate change — it is **equivariant**, meaning it transforms together with the coordinates so that any equation between tensors is automatically coordinate-independent.

To make this concrete, consider changing from one coordinate system $\{x^\mu\}$ to another $\{x^{\mu'}\}$. The two systems are related by a transformation $x^{\mu'} = x^{\mu'}(x^\mu)$, and the Jacobian matrix

$$\Lambda^{\mu'}{}_\nu = \frac{\partial x^{\mu'}}{\partial x^\nu} \quad (13.10)$$

describes how the new coordinates change with the old. The inverse Jacobian is $(\Lambda^{-1})^\mu{}_{\nu'} = \partial x^\mu / \partial x^{\nu'}$.

The chain rule is the definition of a tensor. The chain rule for differentiation is the fundamental reason tensors exist. If $\phi(x)$ is a scalar function, then its gradient in the new coordinates is:

$$\frac{\partial \phi}{\partial x^{\mu'}} = \frac{\partial x^\nu}{\partial x^{\mu'}} \frac{\partial \phi}{\partial x^\nu} = (\Lambda^{-1})^\nu{}_{\mu'} \frac{\partial \phi}{\partial x^\nu}. \quad (13.11)$$

The gradient picks up the inverse Jacobian. Contrast this with how a position displacement dx^μ transforms:

$$dx^{\mu'} = \frac{\partial x^{\mu'}}{\partial x^\nu} dx^\nu = \Lambda^{\mu'}{}_\nu dx^\nu. \quad (13.12)$$

The displacement picks up the Jacobian itself (not the inverse). These two transformation laws are dual to each other, and they define the two species of tensors:

- A **contravariant vector** V^μ transforms with the Jacobian:

$$V^{\mu'} = \Lambda^{\mu'}{}_\nu V^\nu. \quad (13.13)$$

Prototype: position displacement dx^μ , 4-velocity $u^\mu = dx^\mu/d\tau$, 4-momentum p^μ .

- A **covariant vector** (or 1-form) W_μ transforms with the inverse Jacobian:

$$W_{\mu'} = (\Lambda^{-1})^\nu{}_{\mu'} W_\nu. \quad (13.14)$$

Prototype: gradient $\partial_\mu \phi = \partial \phi / \partial x^\mu$, the 4-potential A_μ .

A **tensor of type** (r, s) has r contravariant (upper) indices and s covariant (lower) indices, each transforming by the appropriate Jacobian:

$$T^{\mu'_1 \dots \mu'_r}{}_{\nu'_1 \dots \nu'_s} = \Lambda^{\mu'_1}{}_{\alpha_1} \dots \Lambda^{\mu'_r}{}_{\alpha_r} (\Lambda^{-1})^{\beta_1}{}_{\nu'_1} \dots (\Lambda^{-1})^{\beta_s}{}_{\nu'_s} T^{\alpha_1 \dots \alpha_r}{}_{\beta_1 \dots \beta_s}. \quad (13.15)$$

This is the chain rule applied $r + s$ times.

Why equations between tensors are invariant. If two tensors of the same type are equal in one coordinate system ($A^{\mu\nu} = B^{\mu\nu}$), they are equal in every coordinate system. This follows immediately from the transformation law (13.15): both sides transform identically, so equality is preserved. A physical law written as an equation between tensors is automatically valid in all frames. No frame is preferred; the law does not care which coordinates you use to express it.

The Minkowski metric $\eta_{\mu\nu}$ is a special tensor: it has the same components $\text{diag}(-1, +1, +1, +1)$ in every inertial frame. This is the defining property of Lorentz transformations — they are exactly those Jacobians Λ satisfying $\Lambda^\alpha{}_\mu \Lambda^\beta{}_\nu \eta_{\alpha\beta} = \eta_{\mu\nu}$. The metric is preserved; it is the invariant structure of Minkowski spacetime.

Contracting indices: building scalars from tensors. The Lorentz-invariant interval $ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu$ is formed by contracting a $(0, 2)$ tensor ($\eta_{\mu\nu}$) with two contravariant vectors (dx^μ, dx^ν). The contraction of a covariant and a contravariant index always produces an object with two fewer indices, and if you contract all indices you get a scalar — a number that is the same in every frame. This is the Lorentz-invariant, and it is exactly what the chain rule guarantees: the Jacobians cancel between the covariant and contravariant factors.

Why Maxwell's equations demand tensor language. Before the tensor formulation, Maxwell's equations in 3d vector notation are four separate equations:

$$\nabla \cdot \mathbf{E} = \rho/\epsilon_0, \quad \nabla \times \mathbf{B} - \partial_t \mathbf{E}/c^2 = \mu_0 \mathbf{j}, \quad (13.16)$$

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{E} + \partial_t \mathbf{B} = 0. \quad (13.17)$$

These look like four unrelated laws involving two separate vectors $\mathbf{E}$ and $\mathbf{B}$. Under a Lorentz boost, $\mathbf{E}$ and $\mathbf{B}$ mix:

$$E'_\parallel = E_\parallel, \quad E'_\perp = \gamma(E_\perp + vB_\perp), \quad (13.18)$$

$$B'_\parallel = B_\parallel, \quad B'_\perp = \gamma(B_\perp - vE_\perp/c^2). \quad (13.19)$$

Electric and magnetic fields are not separately Lorentz invariant. They are components of a single object seen from different angles in spacetime — just as the x and y components of a 3d vector mix under rotations.

That single object is the electromagnetic field tensor $F_{\mu\nu}$, a $(0, 2)$ antisymmetric tensor already displayed in (13.9). Its six independent components are precisely the three components of $\mathbf{E}$ and the three components of $\mathbf{B}$. Under a Lorentz boost, $F_{\mu\nu}$ transforms by (13.15) with $r = 0$, $s = 2$, and the mixing of $\mathbf{E}$ and $\mathbf{B}$ is exactly what the Jacobians produce.

All four Maxwell equations collapse into two tensor equations:

$$\partial_\mu F^{\mu\nu} = \mu_0 J^\nu, \quad (13.20)$$

$$\partial_{[\mu} F_{\nu\rho]} = 0. \quad (13.21)$$

Both sides of each equation are tensors of the same type; both equations are therefore automatically valid in every inertial frame. No separate proof of Lorentz invariance is needed: the tensor form carries it for free, via the chain rule.

The current $J^\mu = (\rho c, \mathbf{j})$ is a contravariant 4-vector: charge density and current density are components of one object, just as $\mathbf{E}$ and $\mathbf{B}$ are components of $F_{\mu\nu}$. Equation (13.20) says: the 4-divergence of the field tensor equals the 4-current. This is one equation. It encodes Gauss's law (the $\nu = 0$ component) and Ampère-Maxwell (the $\nu = i$ components) simultaneously. Equation (13.21) — the Bianchi identity $dF = 0$ in index notation — encodes the other two Maxwell equations.

The payoff: tensor language marries time–space and energy–momentum. The tensor formulation does two things that the 3d vector formulation cannot:

1. It makes manifest that **space and time are components of one object**. The position 4-vector $x^\mu = (ct, x, y, z)$ is a single contravariant tensor. The separation between time and space is frame-dependent; different observers split x^μ into time and space components differently. Only $ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu$ is absolute.
2. It makes manifest that **energy and momentum are components of one object**. The 4-momentum $p^\mu = (E/c, \mathbf{p})$ is a single contravariant tensor. The dispersion relation $\eta_{\mu\nu} p^\mu p^\nu = -m^2 c^2$ is a Lorentz-invariant scalar equation involving one tensor. Energy and

momentum are not separately conserved in a Lorentz-invariant sense; only the 4-momentum is.

This is why the tensor language is indispensable for relativistic physics: it forces the coupling of time with space and energy with momentum at the level of notation. You cannot write a Lorentz-invariant equation using only 3d vectors; you are forced to use 4-tensors, and the 4-tensor automatically packages the time and energy components alongside the space and momentum components.

When Dirac tried to write a Lorentz-invariant wave equation, he was compelled by this same logic to write it in terms of $\partial_\mu = (\partial_t/c, \nabla)$ (a covariant 4-vector) and p^μ (a contravariant 4-vector), with the equation relating them through the Minkowski metric. The tensor language made the structure of the problem visible, and the Clifford algebra was the key to solving it.

13.2.3 Energy-momentum: the 4-vector

In Newtonian mechanics, energy and momentum are separate conserved quantities. In special relativity they combine into a single **4-momentum**:

$$p^\mu = (E/c, \mathbf{p}) = m \frac{dx^\mu}{d\tau}, \quad (13.22)$$

where τ is proper time and m is rest mass. The Lorentz-invariant relation:

$$\eta_{\mu\nu} p^\mu p^\nu = -(E/c)^2 + |\mathbf{p}|^2 = -(mc)^2, \quad (13.23)$$

gives the **relativistic energy-momentum relation**:

$$E^2 = (mc^2)^2 + (pc)^2. \quad (13.24)$$

Setting $\mathbf{p} = 0$: $E = mc^2$ (rest energy). For a massless particle: $E = pc$ (photons). For small $p \ll mc$: $E \approx mc^2 + p^2/2m + \dots$ (rest energy plus nonrelativistic kinetic energy).

The coupling of time and space in the Lorentz transformation is the coupling of energy and momentum in the dispersion relation. They are the same phenomenon in position space and momentum space respectively.

13.3 The relativistic wave equation: Klein-Gordon and its failure

13.3.1 From classical to quantum: the substitution rule

Quantum mechanics promotes classical observables to operators. The Schrödinger prescription:

$$E \rightarrow i\hbar \frac{\partial}{\partial t}, \quad \mathbf{p} \rightarrow -i\hbar \nabla. \quad (13.25)$$

For the nonrelativistic energy $E = p^2/2m$, this gives the Schrödinger equation:

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \nabla^2 \psi. \quad (13.26)$$

Now apply the substitution (13.25) to the relativistic relation (13.24): $E^2 = m^2 c^4 + p^2 c^2$ becomes:

$$-\hbar^2 \frac{\partial^2 \psi}{\partial t^2} = (-\hbar^2 c^2 \nabla^2 + m^2 c^4) \psi, \quad (13.27)$$

or in covariant form:

$$\left(\square + \frac{m^2 c^2}{\hbar^2} \right) \psi = 0, \quad \square = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \nabla^2. \quad (13.28)$$

This is the **Klein-Gordon equation**: the natural relativistic wave equation from applying quantum mechanics to the relativistic energy relation.

13.3.2 Why Klein-Gordon fails for a single particle

The Klein-Gordon equation (13.28) has three serious problems as a single-particle wave equation.

Problem 1: No positive-definite probability. The Schrödinger equation conserves the probability $\rho = |\psi|^2 \geq 0$. For the Klein-Gordon equation, the conserved “probability” density is:

$$\rho_{\text{KG}} = \frac{i\hbar}{2mc^2} \left(\psi^* \frac{\partial \psi}{\partial t} - \psi \frac{\partial \psi^*}{\partial t} \right), \quad (13.29)$$

which can be *negative*. A negative probability density is physically meaningless. Probabilities must be non-negative.

Problem 2: Second order in time. The Schrödinger equation is first order in $\partial/\partial t$: specifying $\psi(t=0)$ determines ψ for all t . The Klein-Gordon equation is second order in time: both $\psi(t=0)$ and $\dot{\psi}(t=0)$ must be specified as initial conditions. This is the equation of a field (two initial data per spacetime point), not a single particle (one initial datum per spacetime point: the wavefunction).

Problem 3: Inconsistency with quantum mechanical probability interpretation. This is the deepest problem. The Schrödinger equation is the statement that quantum evolution is unitary: the total probability is conserved by a one-parameter group of unitary transformations. Unitarity requires the time evolution to be a linear, first-order ODE in $\mathbb{C}$ -Hilbert space:

$$i\hbar \frac{\partial |\psi\rangle}{\partial t} = H|\psi\rangle, \quad (13.30)$$

for some self-adjoint Hamiltonian H . The equation must be *first order in time* because:

- The state $|\psi(t)\rangle$ at time t determines the state at all future times (Markov property).
- The evolution $|\psi(t)\rangle = U(t)|\psi(0)\rangle$ must form a group $U(t_1+t_2) = U(t_1)U(t_2)$ (composition).
- A group of unitary operators in Hilbert space is generated by a self-adjoint operator H via $U(t) = e^{-iHt/\hbar}$, which satisfies the first-order equation (13.30).

The Klein-Gordon equation, being second order in time, is incompatible with unitary single-particle quantum mechanics. To have ψ satisfy a second-order equation, you would need to double the state space (include $\dot{\psi}$ as an independent variable), which is no longer the framework of a single quantum particle with a Hilbert space of states.

Summary. A relativistic single-particle wave equation must:

- Be Lorentz invariant (special relativity).
- Be first order in $\partial/\partial t$ (quantum mechanics).
- Conserve a positive-definite probability.

These requirements seem contradictory: special relativity via (13.24) produces a second-order equation. Dirac’s insight was to escape this contradiction by factoring the second-order equation into a first-order one.

13.4 Dirac’s discovery: linearizing the relativistic Hamiltonian

13.4.1 The strategy: factor the operator

Dirac asked: can I find a first-order operator D such that $D^2 = (\text{Klein-Gordon operator})$?

$$D^2 = -\hbar^2 \frac{\partial^2}{\partial t^2} - \hbar^2 c^2 \nabla^2 + m^2 c^4. \quad (13.31)$$

If $D\psi = 0$ implies $D^2\psi = 0$ (i.e., ψ satisfies Klein-Gordon), then a first-order relativistic equation $D\psi = 0$ would do the job.

Try the ansatz: $D = \alpha^\mu p_\mu + \beta mc$, where $p_\mu = (i\hbar\partial_t/c, -i\hbar\nabla)$ and α^μ, β are coefficients to be determined. Then:

$$D^2 = (\alpha^\mu p_\mu + \beta mc)(\alpha^\nu p_\nu + \beta mc) \quad (13.32)$$

$$= \alpha^\mu \alpha^\nu p_\mu p_\nu + (\alpha^\mu \beta + \beta \alpha^\mu) mc p_\mu + \beta^2 m^2 c^2. \quad (13.33)$$

For this to equal $\eta^{\mu\nu} p_\mu p_\nu + m^2 c^2$ (the Klein-Gordon operator), we need:

$$\{\alpha^\mu, \alpha^\nu\} = 2\eta^{\mu\nu} \cdot \mathbf{1}, \quad (13.34)$$

$$\{\alpha^\mu, \beta\} = 0, \quad (13.35)$$

$$\beta^2 = \mathbf{1}, \quad (13.36)$$

where $\{A, B\} = AB + BA$ is the anticommutator and $\mathbf{1}$ is the identity.

13.4.2 Why numbers cannot work: the need for matrices

Can α^μ, β be ordinary numbers? The relation $\{\alpha^\mu, \alpha^\nu\} = 2\eta^{\mu\nu}$ requires, for $\mu = \nu = 0$: $(\alpha^0)^2 = -1$ (since $\eta^{00} = -1$), and for $\mu = \nu = i$: $(\alpha^i)^2 = +1$. But equation (13.35) requires $\alpha^\mu \beta = -\beta \alpha^\mu$: they anticommute. If these were numbers, anticommutativity would imply $\alpha^\mu \beta = -\alpha^\mu \beta$, giving $\alpha^\mu \beta = 0$, contradicting $\alpha^\mu, \beta \neq 0$.

Numbers cannot satisfy these relations. But **matrices** can. The relations (13.34)–(13.36) define the **Clifford algebra** $\text{Cl}(1, 3)$: an algebra of matrices with specific anticommutation relations. The minimal faithful representation is in 4×4 matrices.

Dirac's insight: the wavefunction ψ must therefore be a *vector* (column vector) on which these matrices act, not a scalar. A relativistic single-particle wavefunction is not a complex number at each spacetime point but a vector in $\mathbb{C}^4$ at each spacetime point.

13.4.3 The Dirac matrices and the Dirac equation

The standard representation (Dirac-Pauli basis) uses 4×4 matrices built from the 2×2 Pauli matrices σ_i :

$$\gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \gamma^i = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix}, \quad (13.37)$$

where $\gamma^\mu = (\gamma^0, \gamma^i)$ are the Dirac gamma matrices, satisfying:

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \cdot I_{4 \times 4}. \quad (13.38)$$

This is the Clifford algebra relation. With $\alpha^\mu = (\gamma^0)^{-1} \gamma^\mu$ and absorbing factors, the Dirac equation is:

$$\boxed{(i\hbar\gamma^\mu \partial_\mu - mc)\psi = 0}, \quad (13.39)$$

or in Feynman slash notation $(i\hbar\cancel{\partial} - mc)\psi = 0$. The wavefunction $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)^T$ is a **Dirac spinor**: a 4-component complex vector field.

How Dirac found it. Dirac wrote the equation on a blackboard in 1928 by the following reasoning: “I was playing around with the equation and I suddenly realized that if you take these matrices, the equation would be of first order in the time derivative. And I thought: that’s the equation.” The remarkable thing is that he demanded only two things — first-order in time, Lorentz invariant — and the unique answer forced a 4-component wavefunction, the Clifford algebra, and spin as automatic consequences.

13.5 The Dirac Hamiltonian motor

The Dirac equation $(i\hbar\gamma^\mu\partial_\mu - mc)\psi = 0$ is first order in time. It therefore has the form of a Schrödinger equation $i\hbar\partial_t\psi = H_D\psi$ with a specific Hamiltonian. Isolating the time derivative:

$$i\hbar\frac{\partial\psi}{\partial t} = H_D\psi = (c\boldsymbol{\alpha} \cdot \mathbf{p} + \beta mc^2)\psi, \quad (13.40)$$

where $\boldsymbol{\alpha} = \gamma^0\boldsymbol{\gamma}$ and $\beta = \gamma^0$ are the 4×4 Dirac matrices in the standard representation:

$$\alpha^i = \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}. \quad (13.41)$$

The Hamiltonian H_D is a 4×4 matrix-valued operator acting on the 4-component spinor ψ . In the momentum basis ($\mathbf{p} = \hbar\mathbf{k}$), it is a 4×4 matrix at each momentum:

$$H_D(\mathbf{p}) = \begin{pmatrix} mc^2 I & c\boldsymbol{\sigma} \cdot \mathbf{p} \\ c\boldsymbol{\sigma} \cdot \mathbf{p} & -mc^2 I \end{pmatrix}. \quad (13.42)$$

This is the Dirac Hamiltonian motor at a single momentum $\mathbf{p}$: a 4×4 Hermitian matrix that rotates the 4-component spinor at each tick.

13.5.1 The eigenstructure

The eigenvalues of $H_D(\mathbf{p})$ are found by squaring:

$$H_D(\mathbf{p})^2 = (c^2p^2 + m^2c^4)I_{4 \times 4}, \quad (13.43)$$

where $p = |\mathbf{p}|$. This follows from the Clifford algebra $\{\alpha^i, \alpha^j\} = 2\delta^{ij}$, $\{\alpha^i, \beta\} = 0$, $\beta^2 = I$. Since $H_D^2 = E_p^2 I$ with $E_p = \sqrt{p^2c^2 + m^2c^4}$, the eigenvalues of H_D are:

$$E = \pm E_p = \pm \sqrt{p^2c^2 + m^2c^4}. \quad (13.44)$$

Each sign has a two-fold degeneracy (corresponding to spin up and spin down along any chosen axis), giving four eigenvalues in total: $+E_p$ (twice) and $-E_p$ (twice).

The eigenvectors are the **Dirac spinors** $u^{(s)}(\mathbf{p})$ (positive energy, spin $s = \uparrow, \downarrow$) and $v^{(s)}(\mathbf{p})$ (negative energy, spin $s = \uparrow, \downarrow$). In the standard representation, for $\mathbf{p}$ along the z -axis:

$$u^{(\uparrow)}(\mathbf{p}) = N \begin{pmatrix} 1 \\ 0 \\ pc/(E_p + mc^2) \\ 0 \end{pmatrix}, \quad u^{(\downarrow)}(\mathbf{p}) = N \begin{pmatrix} 0 \\ 1 \\ 0 \\ -pc/(E_p + mc^2) \end{pmatrix}, \quad (13.45)$$

$$v^{(\uparrow)}(\mathbf{p}) = N \begin{pmatrix} pc/(E_p + mc^2) \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad v^{(\downarrow)}(\mathbf{p}) = N \begin{pmatrix} 0 \\ -pc/(E_p + mc^2) \\ 0 \\ 1 \end{pmatrix}, \quad (13.46)$$

where $N = \sqrt{(E_p + mc^2)/2E_p}$ is the normalisation. In the nonrelativistic limit ($p \ll mc$): the positive-energy spinors have large upper components and small lower components (of order v/c); the negative-energy spinors have large lower components and small upper components. In the ultrarelativistic limit ($p \gg mc$): all four components are comparable, and the distinction between “upper” and “lower” dissolves.

The eigenstructure of the Dirac Hamiltonian at each momentum is summarised as:

Eigenvalue	Degeneracy	Spinors	Interpretation
$+E_p$	2	$u^{(\uparrow)}, u^{(\downarrow)}$	positive energy, spin $\uparrow\downarrow$
$-E_p$	2	$v^{(\uparrow)}, v^{(\downarrow)}$	negative energy, spin $\uparrow\downarrow$

The general solution of the Dirac equation is a superposition of plane waves:

$$\psi(x) = \int \frac{d^3p}{(2\pi\hbar)^3} \frac{1}{\sqrt{2E_p}} \sum_{s=\uparrow,\downarrow} \left[b_s(\mathbf{p}) u^{(s)}(\mathbf{p}) e^{-ip\cdot x/\hbar} + d_s^*(\mathbf{p}) v^{(s)}(\mathbf{p}) e^{+ip\cdot x/\hbar} \right], \quad (13.47)$$

where $b_s(\mathbf{p})$ and $d_s^*(\mathbf{p})$ are the coefficients of the positive- and negative-energy solutions respectively.

13.5.2 The problem of negative energy

The negative-energy eigenvalues $E = -E_p$ are the first crisis of relativistic quantum mechanics.

In nonrelativistic quantum mechanics, the harmonic oscillator has a ground state: the energy is bounded below by $E_0 = \frac{1}{2}\hbar\omega$. The ladder operators lower the energy but cannot go below the ground state: $a|0\rangle = 0$. The spectrum has a floor.

The Dirac spectrum has no floor. The negative-energy solutions extend to $E = -\infty$ (as $p \rightarrow \infty$). A positive-energy electron could, in principle, radiate a photon and transition to a negative-energy state, releasing energy $\geq 2mc^2$. Nothing in the single-particle Dirac equation prevents this. And once the electron reaches the negative-energy sea, it can radiate more and cascade to ever-more-negative energies. The system is unstable: there is no ground state.

This instability is catastrophic. A single-particle quantum theory requires a ground state (the lowest-energy eigenstate from which $a|0\rangle = 0$, the state that cannot be lowered further). Without a ground state, the theory has no stable vacuum, no well-defined thermodynamics, no consistent probability interpretation.

Dirac's sea (1930). Dirac proposed a desperate but brilliant solution: all negative-energy states are *already filled* by an infinite sea of electrons, one per state, obeying the Pauli exclusion principle. A positive-energy electron cannot transition to a negative-energy state because all negative-energy states are occupied. The vacuum is not empty; it is a Fermi sea of negative-energy electrons filling every available state down to $E = -\infty$.

A negative-energy electron can be *excited* out of the sea by absorbing a photon of energy $\geq 2mc^2$. The result: a positive-energy electron (visible as an ordinary electron) plus a “hole” in the negative-energy sea. The hole behaves like a particle with positive energy and positive charge: the **positron**, predicted by Dirac in 1931, discovered by Anderson in 1932.

The Dirac sea is conceptually extravagant — an infinite number of invisible particles filling the vacuum — but it makes a correct physical prediction (the positron) and resolves the stability problem (the filled sea prevents transitions to negative energies).

The real resolution: second quantization. The Dirac sea is a scaffolding that is removed by second quantization — the promotion of the Dirac field ψ from a wavefunction to an operator. In the second-quantized theory, the coefficients $b_s(\mathbf{p})$ and $d_s^*(\mathbf{p})$ in the mode expansion (13.47) become *operators*:

$$\hat{\psi}(x) = \int \frac{d^3p}{(2\pi\hbar)^3} \frac{1}{\sqrt{2E_p}} \sum_s \left[\hat{b}_s(\mathbf{p}) u^{(s)}(\mathbf{p}) e^{-ip\cdot x/\hbar} + \hat{d}_s^*(\mathbf{p}) v^{(s)}(\mathbf{p}) e^{+ip\cdot x/\hbar} \right]. \quad (13.48)$$

The crucial reinterpretation: $\hat{d}_s^*(\mathbf{p})$ in the classical solution becomes $\hat{d}_s^\dagger(\mathbf{p})$ — a *creation operator* for an antiparticle (the positron), not a coefficient of a negative-energy solution. The negative-energy solution $v^{(s)}(\mathbf{p}) e^{+ip \cdot x/\hbar}$ does not describe a negative-energy particle; it describes the creation of a positive-energy antiparticle.

The operators satisfy anticommutation relations (because the electron is a fermion):

$$\{\hat{b}_s(\mathbf{p}), \hat{b}_{s'}^\dagger(\mathbf{p}')\} = (2\pi\hbar)^3 \delta_{ss'} \delta^3(\mathbf{p} - \mathbf{p}'), \quad (13.49)$$

$$\{\hat{d}_s(\mathbf{p}), \hat{d}_{s'}^\dagger(\mathbf{p}')\} = (2\pi\hbar)^3 \delta_{ss'} \delta^3(\mathbf{p} - \mathbf{p}'), \quad (13.50)$$

with all other anticommutators vanishing. The anticommutation relations (rather than the commutation relations of the bosonic harmonic oscillator) enforce the Pauli exclusion principle: $(\hat{b}_s^\dagger)^2 = 0$ (you cannot create two electrons in the same state).

13.5.3 The Dirac Hamiltonian motor: second-quantized form

Substituting the field operator (13.48) into the Hamiltonian and normal-ordering (placing all creation operators to the left of all annihilation operators), the second-quantized Dirac Hamiltonian is:

$$H_D = \int \frac{d^3p}{(2\pi\hbar)^3} E_p \sum_s \left[\hat{b}_s^\dagger(\mathbf{p}) \hat{b}_s(\mathbf{p}) + \hat{d}_s^\dagger(\mathbf{p}) \hat{d}_s(\mathbf{p}) \right]. \quad (13.51)$$

This is the Dirac Hamiltonian motor. It is a polynomial of degree 2 in the creation and annihilation operators, exactly as the free harmonic oscillator motor $H_0 = \omega a^\dagger a$ was a polynomial of degree 2. The architecture is the same. The differences are:

1. **Two species per mode.** Each momentum $\mathbf{p}$ and spin s has two operators: $\hat{b}_s^\dagger$ (creates an electron) and $\hat{d}_s^\dagger$ (creates a positron). The motor counts electrons and positrons separately, and both contribute positively to the energy. There are no negative energies.
2. **Anticommutation, not commutation.** The operators satisfy $\{\hat{b}, \hat{b}^\dagger\} = 1$ rather than $[a, a^\dagger] = 1$. The consequence: the occupation number of each mode is 0 or 1 (Pauli exclusion), not 0, 1, 2, 3, ... (Bose-Einstein statistics). The Fock space of a fermionic mode is two-dimensional ($|0\rangle$ and $|1\rangle$), not infinite-dimensional.
3. **Positive-definite energy.** Both terms $\hat{b}^\dagger \hat{b}$ and $\hat{d}^\dagger \hat{d}$ are positive semi-definite, and both are multiplied by the positive energy $E_p > 0$. The negative-energy problem is gone. The motor has a ground state:

$$\hat{b}_s(\mathbf{p}) |0\rangle = 0, \quad \hat{d}_s(\mathbf{p}) |0\rangle = 0 \quad \text{for all } \mathbf{p}, s. \quad (13.52)$$

The vacuum $|0\rangle$ is the state with no electrons and no positrons in any mode. Its energy is zero (after normal ordering). The spectrum is bounded below. The theory is stable.

What second quantization did. The first-quantized Dirac Hamiltonian (13.40) is a 4×4 matrix acting on a 4-component spinor. Its eigenvalues are $\pm E_p$: two positive, two negative. The negative eigenvalues are a crisis.

The second-quantized Dirac Hamiltonian (13.51) is a polynomial of creation and annihilation operators acting on the Fock space. Its spectrum is $\{0, E_p, E_p + E_{p'}, \dots\}$: all energies are sums of positive terms. The negative energies are gone. They have been *reinterpreted*: the negative-energy coefficient $d_s^*(\mathbf{p})$ in the first-quantized theory becomes the positive-energy creation operator $\hat{d}_s^\dagger(\mathbf{p})$ for an antiparticle in the second-quantized theory. The “negative-energy electron” is the “positive-energy positron.” The crisis is resolved not by filling a sea but by changing the language: from wavefunctions to operators, from single particles to Fock space, from a 4×4 matrix to a polynomial motor.

The motor in explicit form. Writing the Hamiltonian (13.51) in terms of number operators $\hat{n}_s^{(e)}(\mathbf{p}) = \hat{b}_s^\dagger(\mathbf{p})\hat{b}_s(\mathbf{p})$ (electron number) and $\hat{n}_s^{(\bar{e})}(\mathbf{p}) = \hat{d}_s^\dagger(\mathbf{p})\hat{d}_s(\mathbf{p})$ (positron number):

$$H_D = \int \frac{d^3p}{(2\pi\hbar)^3} E_p \sum_s \left[\hat{n}_s^{(e)}(\mathbf{p}) + \hat{n}_s^{(\bar{e})}(\mathbf{p}) \right]. \quad (13.53)$$

The motor counts: how many electrons of each momentum and spin, how many positrons of each momentum and spin, and assigns energy E_p to each. The total energy is the sum of all particle energies. The motor is diagonal in the Fock basis. It phase-rotates each Fock sector at its own frequency. It does not create or destroy particles. It is a free motor.

This is the same architecture as the free harmonic oscillator motor $H_0 = \omega a^\dagger a$: a sum of number operators, each multiplied by the energy of that mode. The only structural differences are the anticommutation relations (fermions vs. bosons) and the two-species structure (electrons and positrons per mode vs. one species per mode).

Adding interactions: QED. The free Dirac motor (13.51) is diagonal and creates nothing. The physics of electrons and positrons—pair creation, annihilation, scattering, radiation—requires the interaction with the electromagnetic field. The QED interaction Hamiltonian is:

$$H_{\text{int}}^{\text{QED}} = -e \int d^3x \hat{\psi} \gamma^\mu \hat{\psi} \hat{A}_\mu, \quad (13.54)$$

where $\hat{A}_\mu$ is the photon field (a bosonic operator with its own creation and annihilation operators $\hat{a}_\lambda(\mathbf{k})$, $\hat{a}_\lambda^\dagger(\mathbf{k})$). Expanding all fields in their mode decompositions, the interaction Hamiltonian becomes a polynomial of degree 3 in the creation and annihilation operators: each vertex contains three operators (one $\hat{b}$ or $\hat{b}^\dagger$, one $\hat{d}$ or $\hat{d}^\dagger$, one $\hat{a}$ or $\hat{a}^\dagger$), corresponding to the three lines at each QED vertex in a Feynman diagram.

The full QED motor is:

$$H_{\text{QED}} = \underbrace{H_D + H_\gamma}_{\text{free: degree 2}} + \underbrace{H_{\text{int}}^{\text{QED}}}_{\text{interaction: degree 3}}, \quad (13.55)$$

where $H_\gamma = \int \frac{d^3k}{(2\pi)^3} \omega_k \sum_\lambda \hat{a}_\lambda^\dagger(\mathbf{k}) \hat{a}_\lambda(\mathbf{k})$ is the free photon Hamiltonian. The motor is a polynomial of degree 3 in creation and annihilation operators for three species: electrons ($\hat{b}$), positrons ($\hat{d}$), and photons ($\hat{a}$). The free part is diagonal (phase-rotates each Fock sector). The interaction part is off-diagonal (creates and destroys particles: an electron emits a photon, a photon creates an electron-positron pair, two electrons scatter by exchanging a photon).

The entire content of quantum electrodynamics—the Lamb shift, the anomalous magnetic moment of the electron, pair production, Compton scattering, Bremsstrahlung, vacuum polarisation—is generated by this degree-3 polynomial motor, applied to the Fock state at every tick.

Harmonic oscillator	Free Dirac field	QED
1 species (a , $a^\dagger$)	2 species ($\hat{b}$, $\hat{d}$) per spin	3 species ($\hat{b}$, $\hat{d}$, $\hat{a}$)
Commutators	Anticommutators ($\hat{b}$, $\hat{d}$)	Mixed
$H = \omega a^\dagger a$ (degree 2)	$H = E_p(\hat{b}^\dagger \hat{b} + \hat{d}^\dagger \hat{d})$ (degree 2)	$H = H_{\text{free}} + e \hat{\psi} \gamma^\mu \hat{\psi} \hat{A}_\mu$ (degree 3)
Diagonal (free)	Diagonal (free)	Off-diagonal (interactions)
$n = 0, 1, 2, \dots$ per mode	$n = 0, 1$ per mode (Pauli)	Bosonic + fermionic
1 ground state: $a 0\rangle = 0$	1 vacuum: $\hat{b} 0\rangle = \hat{d} 0\rangle = 0$	Interacting vacuum $ \Omega\rangle$

The moral. The first-quantized Dirac equation is a 4×4 matrix equation whose eigenstructure contains negative energies—a crisis. The second-quantized Dirac Hamiltonian is a polynomial of creation and annihilation operators whose spectrum is positive semi-definite—a resolution. The resolution is not a trick; it is the recognition that the correct language for relativistic quantum mechanics is not wavefunctions but Fock space, not matrices but polynomials of a and $a^\dagger$. The Hamiltonian motor—a polynomial of creation and annihilation operators acting on the Fock space—is the architecture that accommodates both particles and antiparticles, both positive and negative solutions of the Dirac equation, both fermions and bosons, in a single framework with a stable vacuum and a positive-definite energy spectrum.

The game engine is a polynomial motor. This was true for the harmonic oscillator. It is true for the Dirac field. It is true for QED. It is true for the Standard Model. The degree of the polynomial (2 for free theories, 3 for QED, 4 for non-abelian gauge theories and scalar self-interactions) and the species content (which creation and annihilation operators appear, with which commutation or anticommutation relations) determine the physics. But the architecture—a polynomial of a and $a^\dagger$, applied to the Fock state at each tick—is universal.

13.6 Why physicists call it the most beautiful equation

The Dirac equation

$$(i\hbar\gamma^\mu\partial_\mu - mc)\psi = 0 \quad (13.56)$$

is regularly voted by physicists as the most beautiful equation in all of physics. This is not mere aesthetic preference. The beauty is of a specific, arguable kind: the equation achieves an extraordinary compression of content, deriving things that had to be postulated in every earlier framework.

Beauty through inevitability. Dirac imposed just two requirements — Lorentz invariance and linearity in time — that any sensible relativistic quantum mechanics must satisfy. These two requirements, combined with the substitution rule of quantum mechanics, uniquely determine the equation. There is no freedom, no tunable parameter, no arbitrary choice. The gamma matrices, the 4-component spinor, the coupling to the electromagnetic field, the spin of the electron, and the g -factor all follow from the two requirements alone. The equation could not have been otherwise.

Beauty through surprise. Nobody anticipated what the equation would contain. Dirac was looking for a wave equation; he found a theory of spin, antimatter, and the vacuum. The equation “knows” things its creator did not ask it to know:

- *Spin $\frac{1}{2}$:* the 4-component spinor is not an ad hoc addition. The Clifford algebra forces it. Spin is not a property of the electron that happens to be described by the equation; it is a consequence of Lorentz invariance alone.
- *$g = 2$:* the anomalous magnetic moment of the electron, measured to be $g \approx 2$ before Dirac, is derived. Pauli had to postulate it. Dirac proved it. The most precisely measured dimensionless number in physics falls out of an equation that was not designed to produce it.
- *Antimatter:* the negative-energy solutions, an embarrassment when first seen, turned out to describe antiparticles. The positron was predicted in 1931 and discovered in 1932. No other equation in physics has predicted a new form of matter from pure mathematical reasoning.
- *The vacuum is non-trivial:* the Dirac sea, the need for second quantization, and the infinite-dimensional Fock space all reveal that the vacuum is not empty. It is a structured state with fluctuating particle-antiparticle pairs. This, too, was not asked for.

Beauty through unification. The equation unifies concepts that had been separately postulated: special relativity, quantum mechanics, spin, and electromagnetism. It is the first equation in physics that is simultaneously quantum and relativistic, and the two requirements are not in tension — they fit together with mathematical precision.

Beauty through compression. The equation is eight symbols: $(i\hbar\gamma^\mu\partial_\mu - mc)\psi = 0$. These eight symbols, unpacked, contain: the Lorentz group and its spinor representations, the Clifford algebra $\text{Cl}(1,3)$, the prediction of the positron, the necessity of quantum field theory, the spectrum of hydrogen (including fine structure and the Lamb shift when radiative corrections are included), the chirality structure of the Standard Model, and the Atiyah-Singer index theorem via the index of the Dirac operator. The ratio of content to symbols is remarkable.

Dirac's own view. Dirac believed that mathematical beauty was a guide to physical truth. In his 1963 Scientific American article, he wrote: “It seems that if one is working from the point of view of getting beauty in one's equations, and if one has really a sound insight, one is on a sure road to progress.” The Dirac equation is the clearest example of this philosophy in action: the most beautiful equation in physics is also the most correct, in the sense that every consequence it predicts has been confirmed.

The equation stands as evidence that the universe has a preference for mathematical elegance, or at least that the constraints of consistency and symmetry are powerful enough to determine physical reality from very little input. Two requirements — first order, Lorentz invariant — produce an equation that describes all known fermionic matter in the universe.

13.7 Spin emerges automatically: comparing Dirac and Pauli

13.7.1 The Pauli equation: spin by hand

Before Dirac, the quantum-mechanical description of an electron in a magnetic field required the **Pauli equation**, in which spin was introduced by hand:

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix}\phi_\uparrow \\ \phi_\downarrow\end{pmatrix} = \left[\frac{(\mathbf{p} - e\mathbf{A})^2}{2m} - e\phi - \frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B}\right]\begin{pmatrix}\phi_\uparrow \\ \phi_\downarrow\end{pmatrix}. \quad (13.57)$$

The 2-component spinor $(\phi_\uparrow, \phi_\downarrow)$ describes spin-up and spin-down. The term $-\frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B}$ is the magnetic moment coupling, added by hand with the measured value $g = 2$ for the electron. The Pauli equation is *not* derived; it is *postulated* to match experiment.

13.7.2 Coupling Dirac to the electromagnetic field

The minimal coupling rule (reviewed in Chapter 10): replace $\partial_\mu \rightarrow D_\mu = \partial_\mu + \frac{ie}{\hbar}A_\mu$, where A_μ is the electromagnetic 4-potential. The Dirac equation with EM coupling:

$$(i\hbar\gamma^\mu D_\mu - mc)\psi = (i\hbar\gamma^\mu\partial_\mu - e\gamma^\mu A_\mu - mc)\psi = 0. \quad (13.58)$$

13.7.3 Deriving spin and $g = 2$ from the Dirac equation

Taking the non-relativistic limit of (13.58): write $\psi = e^{-imc^2t/\hbar}\begin{pmatrix}\phi \\ \chi\end{pmatrix}$ where ϕ is the large (upper) component and χ is the small (lower) component. In the limit $v \ll c$, the small component satisfies $\chi \approx \frac{\boldsymbol{\sigma} \cdot \mathbf{p}}{2mc}\phi$ and the equation for ϕ reduces to:

$$i\hbar\frac{\partial\phi}{\partial t} = \left[\frac{(\mathbf{p} - e\mathbf{A})^2}{2m} - e\phi - \underbrace{\frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B}}_{\text{spin coupling, } g=2} + \underbrace{\text{spin-orbit, Darwin terms}}_{\text{relativistic corrections}}\right]\phi. \quad (13.59)$$

The magnetic moment term $-\frac{e\hbar}{2m}\boldsymbol{\sigma} \cdot \mathbf{B}$ emerges automatically, with coefficient exactly $1 \times \frac{e\hbar}{m}$, which corresponds to gyromagnetic ratio $g = 2$.

This is the triumph of the Dirac equation. The Pauli equation *postulated* $g = 2$. The Dirac equation *derives* $g = 2$ from first principles: the only requirements were Lorentz invariance and first-order linearity. The fact that the electron has $g \approx 2$ (experimentally measured before Dirac) was a mysterious coincidence in the Pauli framework. In the Dirac framework it is inevitable.

The Pauli versus Dirac comparison:

Feature	Pauli equation	Dirac equation
Spinor components	2	4
Spin coupling	Postulated ($g = 2$ by hand)	Derived ($g = 2$ automatically)
Relativistic	No	Yes
Negative energy	No	Yes
Antimatter	Not predicted	Predicted
Valid for	$v \ll c$	All velocities
EM coupling	Minimal + Pauli term	Minimal coupling only

13.8 The crisis: negative energy states

13.8.1 What negative energy means

The Dirac equation (13.39) for a free particle has plane wave solutions $\psi = u(p)e^{ip \cdot x/\hbar}$ where $u(p)$ satisfies $(\gamma^\mu p_\mu - mc)u = 0$. This equation has four independent solutions for each momentum $\mathbf{p}$: two with positive energy $E = +\sqrt{(pc)^2 + (mc^2)^2} > 0$ and two with negative energy $E = -\sqrt{(pc)^2 + (mc^2)^2} < 0$.

The negative-energy solutions cannot be discarded. They are part of the spectrum and, crucially, the positive and negative-energy sectors are coupled: an electron can emit a photon and transition from positive to negative energy. Since there is no lower bound on the negative energies ($E \rightarrow -\infty$ as $|\mathbf{p}| \rightarrow \infty$), the electron would cascade downward indefinitely, emitting infinite energy. The ground state would not exist.

13.8.2 Why Schrödinger evolution makes it catastrophic

The Schrödinger equation (13.30) is a first-order linear ODE in Hilbert space. Formally:

$$|\psi(t)\rangle = e^{-iHt/\hbar}|\psi(0)\rangle. \quad (13.60)$$

If the Hamiltonian H has no lower bound (negative-energy states extending to $-\infty$), then the spectrum of H is unbounded below. The operator $e^{-iHt/\hbar}$ involves $e^{i|E|t/\hbar}$ for the negative-energy modes, and there is no stable ground state. A particle initially in a positive-energy state will tunnel, via photon emission, into arbitrarily negative-energy states. The transition rate per unit time diverges because there are infinitely many states of lower energy to fall into. The Dirac vacuum — all states empty — is catastrophically unstable.

Heisenberg is said to have called the period 1928–1930 “the saddest time in physics”: Dirac’s beautiful equation, which had derived the electron’s spin and magnetic moment from first principles, seemed to predict that all matter was unstable. Every electron would radiate itself into the void within a fraction of a second.

13.8.3 How the crisis manifests mathematically

In the single-particle Dirac equation, the Hamiltonian is:

$$H_D = c\boldsymbol{\alpha} \cdot \mathbf{p} + \beta mc^2, \quad \boldsymbol{\alpha} = \gamma^0 \boldsymbol{\gamma}, \quad \beta = \gamma^0. \quad (13.61)$$

The spectrum of H_D on $L^2(\mathbb{R}^3) \otimes \mathbb{C}^4$ is $(-\infty, -mc^2] \cup [mc^2, +\infty)$: two bands separated by a gap $2mc^2$, but the negative band is unbounded below. As a single-particle Hamiltonian generating Schrödinger evolution on a Hilbert space, this is a disaster: the evolution $e^{-iH_D t/\hbar}$ is unitary (probability is conserved) but the system has no ground state and any physical state will eventually mix with negative-energy states under any interaction.

13.9 Dirac's hole theory: the first glimpse of the vacuum

13.9.1 The Pauli exclusion principle as the solution

Dirac's resolution (1930) was audacious. He invoked the Pauli exclusion principle: *electrons are fermions, so no two electrons can occupy the same state*. Suppose all negative-energy states are already occupied. Then the Pauli exclusion principle forbids positive-energy electrons from falling into the negative-energy sea: all the lower seats are taken.

The **Dirac sea** is the state with all negative-energy levels filled:

$$|\text{vacuum}\rangle = \prod_{E<0} a_E^\dagger |0\rangle, \quad (13.62)$$

where the product runs over all negative-energy states. This state has infinite negative charge and infinite negative energy (regularized by subtracting them as a reference). The physical vacuum is this filled sea, not the naïve empty vacuum.

Positive-energy electrons: these are additional electrons on top of the sea, behaving normally.

Holes in the sea: if a negative-energy electron is removed (by absorbing a photon of energy $> 2mc^2$), it leaves a **hole** — an absence in the sea. This hole has:

- *Positive energy:* removing a state of energy $E < 0$ leaves the sea with energy $-E > 0$.
- *Positive charge:* the sea has one less negative charge, appearing as charge $+e$.
- *The same mass:* m (since it is the absence of an electron).

A hole in the Dirac sea is a particle with the same mass as the electron but opposite charge: the **positron**.

13.9.2 The prediction of antimatter

In 1931, Dirac predicted the existence of the positron from the hole picture. In 1932, Anderson discovered the positron experimentally in cosmic ray tracks: a particle with the mass of an electron but positive charge, exactly as Dirac predicted.

Pair creation. A photon of energy $\hbar\omega > 2mc^2$ can excite a negative-energy electron to positive energy, creating an electron-positron pair: the photon promotes a sea-electron upward, leaving a hole (positron) and creating a positive-energy electron.

Pair annihilation. A positive-energy electron falling back into a hole in the sea: the electron and positron annihilate, producing photons of total energy $2mc^2$.

13.9.3 Why the hole picture is unsatisfactory

The hole picture works for fermions but has severe problems:

- It relies on the Pauli exclusion principle: it fails completely for bosons (which have no exclusion principle and cannot fill a sea).
- The Dirac sea has infinite charge, energy, and momentum, requiring infinite renormalization.
- It treats particles and holes asymmetrically, while they are physically equivalent.
- It is not a single-particle theory at all: the vacuum is an infinite-body state.

The hole picture was a transitional framework. The correct understanding required abandoning the single-particle interpretation entirely.

13.10 The Fock space: abandoning single-particle quantum mechanics

Why single-particle quantum mechanics must fail

The single-particle framework fails for relativistic quantum mechanics not for one reason but for three, each decisive on its own. Taken together, they leave no room for doubt: a relativistic quantum theory is necessarily a many-body theory from the outset.

Argument 1: The Klein-Gordon equation has a negative probability density. The most natural relativistic wave equation is obtained by promoting the energy-momentum relation $E^2 = \mathbf{p}^2 c^2 + m^2 c^4$ to an operator equation acting on a wavefunction $\phi(x)$:

$$(\partial_\mu \partial^\mu + m^2)\phi = 0, \quad (13.63)$$

the Klein-Gordon equation (setting $\hbar = c = 1$). For a single-particle theory, we need a conserved probability density $\rho \geq 0$. The Klein-Gordon equation has a conserved current:

$$j^\mu = i(\phi^* \partial^\mu \phi - \phi \partial^\mu \phi^*), \quad \partial_\mu j^\mu = 0. \quad (13.64)$$

The time component is:

$$\rho = j^0 = i(\phi^* \partial_t \phi - \phi \partial_t \phi^*). \quad (13.65)$$

For a single-particle Schrödinger equation, $\rho = |\psi|^2 \geq 0$ always. But (13.65) involves the *time derivative* of ϕ , which is an independent initial condition: ρ can be positive or negative depending on the choice of initial data. A probability density that can be negative is not a probability density. There is no fix within the single-particle framework: the problem is structural, not a matter of choice of equation or boundary condition.

Argument 2: The energy spectrum is unbounded below. Both the Klein-Gordon equation (13.63) and the Dirac equation $(i\gamma^\mu \partial_\mu - m)\psi = 0$ have solutions with negative energy $E = -\sqrt{\mathbf{p}^2 + m^2}$ for every momentum $\mathbf{p}$. In a single-particle theory, these negative-energy solutions must be included (the equation has them whether we want them or not), and a particle can emit a photon, drop to a lower energy state, emit another photon, and repeat indefinitely. The energy spectrum is unbounded below: no stable ground state exists. This is not a small technical issue; it is a catastrophic instability that makes the theory physically meaningless.

Dirac's proposed solution — fill all negative-energy states with an infinite sea of electrons, invoking the Pauli exclusion principle to prevent any positive-energy electron from falling in — *immediately abandons the single-particle framework*. The Dirac sea is not a single-particle state; it is a many-body state of infinitely many particles. The very act of invoking the Pauli exclusion principle to stabilise the negative-energy spectrum requires treating the system as a Fermi gas of electrons filling a Fock space of single-particle states. Dirac's fix is Fock space in disguise.

Argument 3: The hole and pair production are already multi-particle phenomena. Dirac's hole — a missing electron in the negative-energy sea — behaves as a positron: a particle

with positive energy, positive charge, and the same mass as the electron. When a photon of sufficient energy $E_\gamma > 2m_e c^2$ is absorbed by the vacuum, an electron is promoted from the negative-energy sea to a positive-energy state, leaving a hole: the process $\gamma \rightarrow e^- + e^+$ (pair creation). The reverse process $e^- + e^+ \rightarrow \gamma$ (pair annihilation) fills a hole. Both processes change the number of observable particles. Pair creation and annihilation are not exotic high-energy phenomena; they are observed at electron-volt scales in positron emission tomography (PET scanning), in cosmic ray showers, and in the Lamb shift (the energy shift of hydrogen levels from virtual electron-positron pairs in the vacuum).

A single-particle wavefunction $\psi(x, t)$ cannot describe a process in which the number of particles changes. The single-particle Hilbert space $\mathcal{H}^{(1)}$ contains states of exactly one particle; it has no room for the vacuum $|0\rangle$ (zero particles) or the two-particle state $|e^-, e^+\rangle$ that are necessary to describe pair creation. This is not a failure of the particular equation chosen; it is a failure of the single-particle framework itself. The correct framework must contain a sector for each particle number simultaneously.

The forced conclusion: Fock space. Each of the three arguments individually requires abandoning the single-particle framework:

Problem	What it shows	Required remedy
Negative probability density	$\rho = j^0$ can be negative in KG equation	j^0 must be reinterpreted as charge density, not probability
Unbounded energy spectrum	Negative-energy solutions exist for all momenta	Infinite-particle Dirac sea; already Fock space
Pair creation/annihilation	Particle number changes; empirically observed	Hilbert space must include all particle-number sectors

The resolution of all three problems is the same: replace the single-particle Hilbert space $\mathcal{H}^{(1)}$ by the Fock space $\mathcal{F} = \bigoplus_n \mathcal{H}^{(n)}$. In the Fock framework: the negative Klein-Gordon “probability density” is reinterpreted as the conserved charge density $j^0 = e(|\phi^+|^2 - |\phi^-|^2)$, the difference between positive-charge and negative-charge particle densities, which can be positive or negative as a charge density should be. The negative-energy states are reinterpreted as positive-energy antiparticle states via the Fock creation operator $d^\dagger(\mathbf{k})$ for the antiparticle (the Feynman-Stueckelberg interpretation): a negative-energy particle moving forward in time is a positive-energy antiparticle moving backward in time. And pair creation is simply the Fock Hamiltonian moving amplitude from the zero-particle sector $\mathcal{H}^{(0)}$ to the two-particle sector $\mathcal{H}^{(2)}$, exactly as any other interaction moves amplitude between sectors. The Dirac sea disappears: it was an artefact of trying to solve a many-body problem within a single-particle framework. In the Fock space, the vacuum $|0\rangle$ is simply the lowest-energy state of the Hamiltonian, with no infinite sea of filled states required.

13.10.1 Fock space: the space of all particle numbers

The **Fock space** is the direct sum of n -particle Hilbert spaces:

$$\mathcal{F} = \bigoplus_{n=0}^{\infty} \mathcal{H}_n, \quad (13.66)$$

where $\mathcal{H}_0 = \mathbb{C}$ (the vacuum), $\mathcal{H}_1 = L^2(\mathbb{R}^3) \otimes \mathbb{C}^4$ (one-particle states), $\mathcal{H}_2$ is the two-particle space (antisymmetrized for fermions), etc.

Creation and annihilation operators.

- $a_s^\dagger(\mathbf{p})$: creates a positive-energy electron with momentum $\mathbf{p}$ and spin s .
- $a_s(\mathbf{p})$: annihilates such an electron.
- $b_s^\dagger(\mathbf{p})$: creates a positron (antiparticle) with momentum $\mathbf{p}$ and spin s .
- $b_s(\mathbf{p})$: annihilates such a positron.

These operators satisfy the canonical anticommutation relations (CAR):

$$\{a_r(\mathbf{p}), a_s^\dagger(\mathbf{q})\} = \delta_{rs} \delta^3(\mathbf{p} - \mathbf{q}), \quad (13.67)$$

$$\{b_r(\mathbf{p}), b_s^\dagger(\mathbf{q})\} = \delta_{rs} \delta^3(\mathbf{p} - \mathbf{q}), \quad (13.68)$$

with all other anticommutators zero.

The quantum field:

$$\hat{\psi}(x) = \int \frac{d^3p}{(2\pi)^3} \sum_s \left[u_s(\mathbf{p}) a_s(\mathbf{p}) e^{ip \cdot x/\hbar} + v_s(\mathbf{p}) b_s^\dagger(\mathbf{p}) e^{-ip \cdot x/\hbar} \right], \quad (13.69)$$

is now an operator-valued distribution: it annihilates electrons and creates positrons at each spacetime point. This is the Dirac quantum field.

13.11 Pauli's reconceptualization: the Dirac equation as a classical field

13.11.1 Field theory, not wave mechanics

Pauli (and Jordan, and Heisenberg) in the early 1930s provided the correct conceptual framework: the Dirac equation should be treated not as a single-particle Schrödinger equation but as a **classical field equation**, analogous to Maxwell's equations. The field $\psi(x)$ is a spinor-valued field on spacetime. It is then *quantized* using the Fock space framework.

The two-step process:

1. **Classical Dirac field:** $\psi(x)$ is a spinor-valued function on spacetime (like $A_\mu(x)$ for the EM field). The Dirac equation $(i\hbar \not{\partial} - mc)\psi = 0$ is its equation of motion.
2. **Canonical quantization:** promote $\psi(x)$ to an operator $\hat{\psi}(x)$ on Fock space satisfying the canonical anticommutation relations (fermion statistics requires anticommutation, not commutation).

13.11.2 The action for the classical Dirac field

The Dirac equation (13.39) is the Euler-Lagrange equation for the action:

$$S[\psi, \bar{\psi}] = \int d^4x \bar{\psi}(x) (i\hbar \gamma^\mu \partial_\mu - mc) \psi(x), \quad (13.70)$$

where $\bar{\psi} = \psi^\dagger \gamma^0$ is the Dirac conjugate. Varying with respect to $\bar{\psi}$ gives the Dirac equation; varying with respect to ψ gives the conjugate equation.

The action (36.24) is Lorentz invariant (because $\bar{\psi}\psi$ is a Lorentz scalar and the Dirac operator transforms correctly) and gauge-invariant under $\psi \rightarrow e^{i\Lambda}\psi$, $A_\mu \rightarrow A_\mu + \partial_\mu \Lambda$.

13.11.3 Second quantization

In the path integral approach, the quantum theory is defined by:

$$Z = \int \mathcal{D}\psi \mathcal{D}\bar{\psi} e^{iS[\psi, \bar{\psi}]/\hbar}, \quad (13.71)$$

where the integration is over anticommuting (Grassmann) fields. This is the quantum field theory of electrons and positrons: quantum electrodynamics (QED) when coupled to the photon field.

The resolution of the negative-energy problem. In the field theory language, the Dirac equation has no “negative energy” problem. The Hamiltonian of the quantized Dirac field:

$$\hat{H} = \int \frac{d^3p}{(2\pi)^3} \sum_s E_{\mathbf{p}} \left[a_s^\dagger(\mathbf{p}) a_s(\mathbf{p}) + b_s^\dagger(\mathbf{p}) b_s(\mathbf{p}) \right] \quad (13.72)$$

(after normal ordering, subtracting the infinite vacuum energy) is *positive definite*: $\hat{H} \geq 0$. The negative-energy modes of the classical field, after second quantization with anticommutation relations, become the positive-energy positron creation operators $b^\dagger$. The “energy” of a classical negative-frequency mode becomes the energy of creating a positron, which is positive.

13.12 What Dirac actually discovered

Dirac set out to find a first-order relativistic wave equation. He found one. But the equation he wrote down contained far more than he was looking for. The four-component wavefunction ψ_A , which appeared as a technical necessity to satisfy the Clifford algebra, turned out to be something profound: a new mathematical object called a *spinor*, which transforms under Lorentz transformations in a way that had never been seen before. The Dirac equation is the first physically realised example of an equivariant representation of the Lorentz group.

The wavefunction transforms when you change frame

In non-relativistic quantum mechanics, the wavefunction $\psi(\mathbf{x}, t)$ is a complex scalar: when you change coordinates, only the argument transforms. The value at each point is a single complex number that does not rotate. For the Klein-Gordon scalar field, the same is true: $\phi'(x) = \phi(\Lambda^{-1}x)$. The spin-0 representation of the Lorentz group is trivial — it is just the number 1.

For the Dirac field, something new happens. The four components $\psi_A(x)$ ($A = 1, 2, 3, 4$) are not independent scalars: they are linked by the Lorentz symmetry. When we change inertial frame by a Lorentz transformation Λ , the field must transform as:

$$\psi'_A(x) = S(\Lambda)^B_A \psi_B(\Lambda^{-1}x), \quad (13.73)$$

where $S(\Lambda)$ is a 4×4 matrix that depends on Λ . The argument $\Lambda^{-1}x$ is the same as for the scalar: the point transforms. But now the *value* also transforms, by the matrix $S(\Lambda)$ acting on the spinor index A . This is the equivariance condition: the field at the transformed point in the new frame equals $S(\Lambda)$ times the field at the original point in the old frame.

The assignment $\Lambda \mapsto S(\Lambda)$ must be a group homomorphism:

$$S(\Lambda_2\Lambda_1) = S(\Lambda_2)S(\Lambda_1), \quad (13.74)$$

so that applying two successive Lorentz transformations gives the same result as applying their composition. This is the definition of a *representation* of the Lorentz group. The content of Dirac’s discovery is that there exists a non-trivial 4×4 matrix representation $S(\Lambda)$ of the Lorentz group, and that the corresponding equivariant field — the Dirac spinor — describes a spin- $\frac{1}{2}$ particle.

The spinor representation from the Clifford algebra

How is $S(\Lambda)$ constructed? The answer comes directly from the Clifford algebra (13.38). For an infinitesimal Lorentz transformation parametrised by the antisymmetric tensor $\omega_{\mu\nu}$ (six parameters: three boosts and three rotations):

$$\Lambda^\mu{}_\nu = \delta^\mu{}_\nu + \omega^\mu{}_\nu + O(\omega^2), \quad (13.75)$$

the corresponding spinor transformation matrix is:

$$S(\Lambda) = I + \frac{i}{4} \omega_{\mu\nu} \sigma^{\mu\nu} + O(\omega^2), \quad \sigma^{\mu\nu} = \frac{i}{2} [\gamma^\mu, \gamma^\nu]. \quad (13.76)$$

The generators of the spinor representation are the six matrices $\sigma^{\mu\nu} = \frac{i}{2} [\gamma^\mu, \gamma^\nu]$, built entirely from the gamma matrices that appeared in the Clifford algebra. The finite transformation is the matrix exponential:

$$S(\Lambda) = \exp\left(\frac{i}{4} \omega_{\mu\nu} \sigma^{\mu\nu}\right). \quad (13.77)$$

This is a 4×4 matrix for every Lorentz transformation Λ . Crucially, $S(\Lambda)$ is *not* the same as Λ itself: the Dirac spinor does not transform as a four-vector. It transforms in a completely different representation of the Lorentz group, one that has no classical analogue.

The key difference: 4π periodicity. For a spatial rotation by angle θ about the z -axis, $\Lambda_{\text{rot}}(\theta)$ is a 4×4 matrix with $\Lambda(2\pi) = I$ (one full rotation returns to the identity). The corresponding spinor transformation is:

$$S(\Lambda_{\text{rot}}(\theta)) = e^{i\theta\sigma^{12}/2} = I \cos \frac{\theta}{2} + i\sigma^{12} \sin \frac{\theta}{2}. \quad (13.78)$$

At $\theta = 2\pi$: $S(\Lambda_{\text{rot}}(2\pi)) = -I$. A full spatial rotation takes the Dirac spinor to *minus itself*. Only at $\theta = 4\pi$ does S return to $+I$. This 4π periodicity is the mathematical signature of the spinor: it transforms in the *double cover* $SL(2, \mathbb{C})$ of the Lorentz group $SO^+(1, 3)$, not in the Lorentz group itself:

$$SL(2, \mathbb{C}) \xrightarrow{2:1} SO^+(1, 3). \quad (13.79)$$

Two distinct elements $\pm S \in SL(2, \mathbb{C})$ correspond to the same Lorentz transformation $\Lambda \in SO^+(1, 3)$. The Dirac spinor “sees” the double cover, not the group itself. This is why fermions acquire a minus sign under 2π rotation and why the spin-statistics theorem links half-integer spin to fermionic anticommutation.

Decomposing the Dirac spinor: Weyl spinors

The 4×4 spinor representation (13.77) is reducible: it decomposes into two irreducible 2×2 representations under the Lorentz group. To see this, introduce the chiral projectors:

$$P_L = \frac{1 - \gamma^5}{2}, \quad P_R = \frac{1 + \gamma^5}{2}, \quad \gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3. \quad (13.80)$$

The Dirac spinor decomposes as:

$$\psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix}, \quad \psi_L = P_L \psi \in \mathbb{C}^2, \quad \psi_R = P_R \psi \in \mathbb{C}^2, \quad (13.81)$$

where ψ_L is the *left-handed Weyl spinor* (2 components, index $s \in \{1, 2\}$) and ψ_R is the *right-handed Weyl spinor* (2 components, index $\dot{s} \in \{1, 2\}$). Under a Lorentz transformation:

$$\psi_L \mapsto L \psi_L, \quad \psi_R \mapsto \bar{L} \psi_R, \quad L \in SL(2, \mathbb{C}), \quad (13.82)$$

where L and $\bar{L}$ are related by complex conjugation but are not the same matrix (they are not equivalent representations for $L \notin SU(2)$, i.e., for boosts). The Lorentz representation content of the Dirac spinor is therefore:

$$\text{Dirac spinor} = \underbrace{\left(\frac{1}{2}, 0\right)}_{\psi_L} \oplus \underbrace{\left(0, \frac{1}{2}\right)}_{\psi_R}, \quad (13.83)$$

in the (j_L, j_R) classification. The left-handed Weyl spinor $(\frac{1}{2}, 0)$ transforms under L but not $\bar{L}$; the right-handed Weyl spinor $(0, \frac{1}{2})$ transforms under $\bar{L}$ but not L . These are the two irreducible spinor representations, and the Dirac spinor is their direct sum.

The mass term mixes chiralities. The mass term in the Dirac equation:

$$m\bar{\psi}\psi = m(\psi_L^\dagger\psi_R + \psi_R^\dagger\psi_L) \quad (13.84)$$

couples ψ_L and ψ_R : a massive Dirac fermion is neither purely left-handed nor purely right-handed. A massless fermion ($m = 0$) decouples: the left and right-handed Weyl equations $i\sigma^\mu\partial_\mu\psi_L = 0$ and $i\bar{\sigma}^\mu\partial_\mu\psi_R = 0$ are independent. The Standard Model exploits this: the weak interaction couples only to ψ_L , and it is the Higgs Yukawa coupling that mixes ψ_L and ψ_R to give fermion masses after the Higgs condensate develops. The left-right asymmetry of the Standard Model is a direct consequence of the irreducibility of the Weyl representations (13.83).

What the Dirac equation tells the Lorentz group

The Dirac equation (13.39) can now be understood as an equivariance condition: it is the equation that a field must satisfy to transform consistently under (13.73). More precisely, it is the equation of motion that is Lorentz invariant when ψ transforms in the $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation. The Lorentz invariance of the Dirac equation:

$$(i\gamma^\mu\partial_\mu - m)\psi = 0 \xrightarrow{\Lambda} S(\Lambda)(i\gamma^\mu\partial_\mu - m)S(\Lambda)^{-1} \cdot S(\Lambda)\psi(\Lambda^{-1}x) = 0 \quad (13.85)$$

holds because the gamma matrices satisfy:

$$S(\Lambda)\gamma^\mu S(\Lambda)^{-1} = \Lambda^\mu{}_\nu\gamma^\nu, \quad (13.86)$$

which follows directly from the definition (13.76) of $S(\Lambda)$ in terms of the Clifford algebra. The gamma matrices are *intertwining operators*: they transform the vector index μ on the left into spinor indices on the right, bridging the vector representation $(\frac{1}{2}, \frac{1}{2})$ and the spinor representation $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$.

This is the deep content of Dirac's discovery, stated in the language of representation theory: the Clifford algebra relation $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ is not merely an algebraic trick for factoring the Klein-Gordon operator. It is the statement that the gamma matrices provide an equivariant map between the vector representation and the spinor representation of the Lorentz group. The Dirac equation is the unique first-order Lorentz-invariant equation for a field in the $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation. Demanding first-order in time and Lorentz invariance, as Dirac did, forced the spinor representation into existence.

Foreshadowing the full Lorentz representation classification

Dirac's spinor is the first entry in a systematic classification of all fields by their Lorentz representation. The pattern is now clear:

Field	Rep (j_L, j_R)	Components	Transformation	Physical example
Scalar	$(0, 0)$	1	$\phi'(x) = \phi(\Lambda^{-1}x)$	Higgs, pion
Left Weyl spinor	$(\frac{1}{2}, 0)$	2	$\psi_L \mapsto L\psi_L$	Left-handed quark, lepton
Right Weyl spinor	$(0, \frac{1}{2})$	2	$\psi_R \mapsto \bar{L}\psi_R$	Right-handed quark, lepton
Dirac spinor	$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	4	$\psi \mapsto S(\Lambda)\psi(\Lambda^{-1}x)$	Electron, quarks
Vector	$(\frac{1}{2}, \frac{1}{2})$	4	$A'_\mu(x) \quad \quad \quad = \quad \quad \quad \Lambda_\mu{}^\nu A_\nu(\Lambda^{-1}x)$	Photon, $W^\pm$, Z^0 , gluon

Every field in the Standard Model is one of these representations, or a direct sum of them. The Fock basis vectors for each field carry the corresponding index: scalars carry no spinor index, Weyl spinors carry index $s \in \{1, 2\}$, Dirac spinors carry index $A \in \{1, 2, 3, 4\}$, and vector fields carry index $\mu \in \{0, 1, 2, 3\}$. The monosemantic Fock basis vector:

$$|n; (\mathbf{k}_1, A_1, \alpha_1), \dots, (\mathbf{k}_n, A_n, \alpha_n)\rangle \quad (13.87)$$

carries the spinor index A_j (from the Dirac representation) and the colour index α_j (from the gauge group representation) on each particle. The Hamiltonian acts on these indices via the gamma matrices and the gauge group generators, all of which are determined by the equivariance conditions. What Dirac discovered in 1928 — that Lorentz invariance forces a new kind of index on the wavefunction, transforming in a representation that does not exist for classical objects — is the foundation of the entire index structure of the Standard Model. Every subsequent enrichment of the Fock basis (adding colour indices, isospin indices, species labels) follows the same logic: Lorentz invariance or gauge invariance forces a new index, the new index transforms in a specific representation, and the representation determines how the Hamiltonian acts on the corresponding Fock basis vectors.

14 Quantizing the Electromagnetic Field

The Dirac equation describes a relativistic electron, but the electron interacts with light—and light itself is a quantum object. Dirac’s quantization of the electromagnetic field (1927) was the first quantum field theory: the first time a classical *field* was promoted to a quantum operator. The result introduces photons as the quanta of the EM field, builds the bosonic Fock space, and explains how the classical electromagnetic wave that Maxwell described emerges as a special quantum state.

14.1 An irony in the history of quantum field theory

In 1927—a full year before he wrote the Dirac equation—Dirac published “The Quantum Theory of the Emission and Absorption of Radiation.” He decomposed the electromagnetic field into Fourier modes, associated a quantum harmonic oscillator to each mode, and introduced the creation operator $a_{\mathbf{k}}^\dagger$ and annihilation operator $a_{\mathbf{k}}$ with $[a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta^3(\mathbf{k} - \mathbf{k}')$. This was the birth of quantum field theory. The recipe:

1. Start with a classical field (A_μ).
2. Expand in modes (Fourier decomposition).
3. Promote each mode to a quantum harmonic oscillator.
4. The Fock space of all occupation numbers is the Hilbert space.

This canonical quantization recipe became the template for all of QFT and the Standard Model: every fundamental particle is a quantum of an underlying field.

The irony: having invented the recipe for the EM field (a boson), Dirac then wrote his electron equation (1928) and did *not* apply his own recipe to it. He treated the Dirac equation as a single-particle wave equation, not a field to be quantized. The negative-energy crisis, the hole theory, and all the conceptual difficulties arose precisely because Dirac did not apply his own quantization recipe to the fermion field.

It was Pauli (with Jordan, Wigner, and Heisenberg in the early 1930s) who applied Dirac’s recipe to the Dirac equation: treat it as a classical field equation, then quantize with *anticommutation* relations $\{a, a^\dagger\} = 1$ (enforcing the Pauli exclusion principle). The Fock space then contains states of definite electron and positron number, the negative-energy problem dissolves, and all particles and antiparticles are quanta of the same underlying field.

14.2 Classical electromagnetism in tensor language

The electromagnetic field is described by the 4-potential $A^\mu = (\phi/c, \mathbf{A})$, with gauge invariance $A_\mu \rightarrow A_\mu + \partial_\mu \Lambda$. The field strength tensor:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix} \quad (14.1)$$

is gauge-invariant ($F_{\mu\nu}[A + \partial\Lambda] = F_{\mu\nu}$, because $\partial_\mu \partial_\nu - \partial_\nu \partial_\mu = 0$) and Lorentz-covariant (it transforms as a $(0, 2)$ tensor).

Maxwell’s four equations reduce to two tensor equations:

$$\partial_\mu F^{\mu\nu} = \mu_0 J^\nu \quad (\text{Gauss} + \text{Ampère}), \quad (14.2)$$

$$\partial_{[\mu} F_{\nu\rho]} = 0 \quad (\text{Faraday} + \text{no monopoles}). \quad (14.3)$$

The second is the Bianchi identity $dF = d(dA) = 0$, automatic from $F = dA$. Current conservation $\partial_\nu J^\nu = 0$ follows from the antisymmetry of $F^{\mu\nu}$.

The Maxwell Lagrangian density:

$$\mathcal{L}_{\text{EM}} = -\frac{1}{4\mu_0} F_{\mu\nu} F^{\mu\nu} + A_\mu J^\mu = \frac{\epsilon_0}{2} (|\mathbf{E}|^2 - c^2 |\mathbf{B}|^2) + A_\mu J^\mu. \quad (14.4)$$

14.3 The electromagnetic Hamiltonian motor

The Schrödinger equation requires a Hamiltonian, not a Lagrangian. The Legendre transform:

Canonical momentum. $\pi^i = \partial\mathcal{L}/\partial(\partial_0 A_i) = -\epsilon_0 E_i$: the electric field is the “momentum” of the vector potential. The momentum conjugate to A_0 vanishes: $\pi^0 = 0$ (because $\mathcal{L}$ does not contain $\dot{A}_0$). This signals gauge freedom: A_0 is a Lagrange multiplier enforcing Gauss’s law, not a dynamical variable.

Hamiltonian density. In Coulomb gauge ($A_0 = 0, \nabla \cdot \mathbf{A} = 0$):

$$\mathcal{H} = \frac{\epsilon_0}{2} (|\mathbf{E}|^2 + c^2 |\mathbf{B}|^2) = \frac{1}{2} (|\boldsymbol{\pi}|^2/\epsilon_0 + \epsilon_0 c^2 |\nabla \times \mathbf{A}|^2). \quad (14.5)$$

Electric energy ($\propto |\mathbf{E}|^2$, the “kinetic” term) plus magnetic energy ($\propto |\mathbf{B}|^2$, the “potential” term). This has the same form as the harmonic oscillator $H = p^2/2m + m\omega^2 x^2/2$: the electric field is the momentum, the vector potential is the coordinate.

Gauge freedom and physical degrees of freedom. The 4-potential A_μ has 4 components, but the gauge freedom $A_\mu \rightarrow A_\mu + \partial_\mu \Lambda$ removes one, and the constraint $\nabla \cdot \mathbf{A} = 0$ (Coulomb gauge) removes another. The physical degrees of freedom are the two transverse polarisations of the vector potential. For each wavevector $\mathbf{k}$, only the two components perpendicular to $\mathbf{k}$ are physical. This is why there are two polarisation vectors $\boldsymbol{\epsilon}_\lambda(\mathbf{k})$, $\lambda = 1, 2$, and the sum over modes is $\sum_{\mathbf{k}, \lambda}$ with $\lambda \in \{1, 2\}$.

14.4 The classical EM field as coupled oscillators

In Coulomb gauge, the wave equation $\nabla^2 \mathbf{A} - \ddot{\mathbf{A}}/c^2 = 0$ with mode expansion

$$\mathbf{A}(\mathbf{x}, t) = \sum_{\mathbf{k}, \lambda} \sqrt{\frac{\hbar}{2\epsilon_0 \omega_{\mathbf{k}} V}} \boldsymbol{\epsilon}_\lambda(\mathbf{k}) [\alpha_{\mathbf{k}\lambda}(t) e^{i\mathbf{k} \cdot \mathbf{x}} + \text{c.c.}] \quad (14.6)$$

gives $\ddot{\alpha}_{\mathbf{k}\lambda} + \omega_{\mathbf{k}}^2 \alpha_{\mathbf{k}\lambda} = 0$ for each mode ($\omega_{\mathbf{k}} = c|\mathbf{k}|$): a harmonic oscillator. The classical EM field is an infinite collection of independent harmonic oscillators, one per mode $(\mathbf{k}, \lambda)$, with total energy $H_{\text{cl}} = \sum_{\mathbf{k}, \lambda} \hbar \omega_{\mathbf{k}} |\alpha_{\mathbf{k}\lambda}|^2$.

14.5 Quantization: each mode becomes a quantum oscillator

Replace each classical amplitude $\alpha_{\mathbf{k}\lambda}$ by an annihilation operator $a_{\mathbf{k}\lambda}$:

$$[a_{\mathbf{k}\lambda}, a_{\mathbf{k}'\lambda'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'} \delta_{\lambda\lambda'}, \quad [a, a] = [a^\dagger, a^\dagger] = 0. \quad (14.7)$$

The vector potential becomes a quantum field operator:

$$\hat{\mathbf{A}}(\mathbf{x}) = \sum_{\mathbf{k}, \lambda} \sqrt{\frac{\hbar}{2\epsilon_0 \omega_{\mathbf{k}} V}} \boldsymbol{\epsilon}_\lambda(\mathbf{k}) [a_{\mathbf{k}\lambda} e^{i\mathbf{k} \cdot \mathbf{x}} + a_{\mathbf{k}\lambda}^\dagger e^{-i\mathbf{k} \cdot \mathbf{x}}]. \quad (14.8)$$

The Hamiltonian, after normal ordering:

$$H = \sum_{\mathbf{k}, \lambda} \hbar \omega_{\mathbf{k}} \hat{n}_{\mathbf{k}\lambda}, \quad \hat{n}_{\mathbf{k}\lambda} = a_{\mathbf{k}\lambda}^\dagger a_{\mathbf{k}\lambda}. \quad (14.9)$$

Energy = sum of (photon number $\times$ photon energy) per mode.

14.6 The Fock space of photons

The Hilbert space is the bosonic Fock space $\mathcal{F}_{\text{photon}} = \bigoplus_{n=0}^{\infty} \mathcal{H}_{\text{sym}}^{(n)}$. The vacuum: $a_{\mathbf{k}\lambda} |0\rangle = 0$ for all $\mathbf{k}, \lambda$; $H |0\rangle = 0$. The one-photon state: $|1_{\mathbf{k}\lambda}\rangle = a_{\mathbf{k}\lambda}^\dagger |0\rangle$ with $H |1_{\mathbf{k}\lambda}\rangle = \hbar\omega_{\mathbf{k}} |1_{\mathbf{k}\lambda}\rangle$. A general n -photon state:

$$|n_{\mathbf{k}_1\lambda_1}, \dots\rangle = \prod_j \frac{(a_{\mathbf{k}_j\lambda_j}^\dagger)^{n_j}}{\sqrt{n_j!}} |0\rangle. \quad (14.10)$$

The general Fock hidden state $h_t = \sum c_{\{n\}}(t) |\{n_{\mathbf{k}\lambda}\}\rangle$ evolves by phase-rotating each coefficient at the total photon energy: $c_{\{n\}}(t) = c_{\{n\}}(0) \exp(-it \sum_{\mathbf{k}\lambda} n_{\mathbf{k}\lambda} \omega_{\mathbf{k}})$.

14.7 The Born rule for photons: projected superpositions in Fock space

The Born protocol applies to the photon Fock space exactly as to any Hilbert space. What makes the photon case rich is the size of the eigensubspaces: measuring one mode's photon number leaves all other modes unresolved.

Level 1: complete measurement. Measure all number operators $\{\hat{n}_{\mathbf{k}\lambda}\}$ simultaneously. Each eigenvalue is non-degenerate: a single Fock basis vector. Born probability $P(\{n\}) = |c_{\{n\}}|^2$. Full information, full trimming.

Level 2: measure one mode. Measure $\hat{n}_{\mathbf{k}_0\lambda_0}$ only. The eigensubspace for eigenvalue n_0 contains all Fock states with n_0 photons in mode $(\mathbf{k}_0, \lambda_0)$ and arbitrary photon numbers in all other modes. The degeneracy is $\prod_{\mathbf{k}\neq\mathbf{k}_0} (n_{\mathbf{k}\lambda} + 1)$: enormous. The projected superposition is a rich superposition over the unresolved modes. Measuring one mode gives partial information; the rest of the photon field is untouched.

Level 3: measure total photon number. Measure $\hat{N} = \sum_{\mathbf{k}\lambda} \hat{n}_{\mathbf{k}\lambda}$. The eigensubspace for total number N contains all ways of distributing N photons across all modes—degeneracy growing combinatorially with the number of modes. The projected superposition preserves all correlations between modes that share the same total count.

Level 4: measure the field value. Measure $\hat{\mathbf{A}}(\mathbf{x})$ at a point. The field operator mixes all modes (it is a sum over $\mathbf{k}$), and the eigenstates are coherent-state-like superpositions over all photon numbers. A definite field value is *maximally indefinite* in photon number.

These four levels are the degeneracy hierarchy applied to photons: from complete resolution (single Fock basis vector, $d = 1$) to field measurement (superposition over all photon numbers, $d \rightarrow \infty$).

14.8 Coherent states: the bridge to classical EM

The coherent state $|\alpha\rangle$ ($\alpha \in \mathbb{C}$), defined by $a|\alpha\rangle = \alpha|\alpha\rangle$, has the Fock decomposition:

$$|\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle. \quad (14.11)$$

Born distribution over photon number: $P(n) = e^{-|\alpha|^2} |\alpha|^{2n} / n!$ (Poisson, mean $\bar{n} = |\alpha|^2$, relative uncertainty $1/|\alpha| \rightarrow 0$ for $|\alpha| \gg 1$).

Under the free Hamiltonian: $e^{-iHt/\hbar} |\alpha\rangle = |\alpha e^{-i\omega t}\rangle$ —the amplitude rotates in the complex plane at the classical frequency, but the Poisson shape is preserved. The coherent state is a “moving fixed point” of the free Schrödinger RNN.

The expectation value of the field operator in a multi-mode coherent state $\bigotimes_{\mathbf{k},\lambda} |\alpha_{\mathbf{k}\lambda}\rangle$ reproduces the classical vector potential (14.6) exactly. The quantum expectation value is the classical field.

14.9 The emergence of the classical EM field

The classical electromagnetic wave emerges from the Fock hidden state through the Born layer with a highly degenerate observable.

The classical field is a coherent state:

$$|\text{classical}\rangle = \bigotimes_{\mathbf{k},\lambda} |\alpha_{\mathbf{k}\lambda}\rangle. \quad (14.12)$$

The Born distribution for the field observable $\hat{\mathbf{A}}(\mathbf{x})$ is a narrow Gaussian centred on the classical value, with variance $\sim 1/|\alpha|^2$. For a classical wave with $|\alpha|^2 \sim 10^{10}$ photons per mode (a laser beam): the Born distribution is near-delta, sampling is deterministic, trimming is inert. The observer sees a definite classical field.

The Born distribution for the number observable $\hat{n}$ is Poisson with mean $|\alpha|^2$: the photon number is uncertain, but the relative uncertainty $1/|\alpha| \rightarrow 0$. The complementarity: a coherent state has a definite field value (near-delta Born distribution for $\hat{\mathbf{A}}$) and an indefinite photon number (Poisson for $\hat{n}$). A number state $|n\rangle$ has a definite photon number and zero field expectation. The same Fock hidden state, read in different Born-layer languages, gives different physics.

State	$P(n)$	$\langle \hat{\mathbf{A}} \rangle$	Character
$ 0\rangle$ (vacuum)	$P(0) = 1$	0	Empty, quiet
$ n_0\rangle$ (number)	$P(n_0) = 1$	0	Definite quanta, no field
$ \alpha\rangle$ (coherent)	Poisson	Classical wave	Classical EM

14.10 The photon vacuum

The free vacuum $|0\rangle$ is the ground state of the free Hamiltonian: $a_{\mathbf{k}\lambda}|0\rangle = 0$ for all modes, $H|0\rangle = 0$. It is a single, definite Fock vector—genuinely empty, no superposition over photon numbers.

The Born distribution for the number operator is exactly delta: $P(n = 0) = 1$ for every mode. But the Born distribution for the field operator $\hat{\mathbf{A}}(\mathbf{x})$ has nonzero variance: $\langle 0|\hat{\mathbf{A}}^2|0\rangle = \sum_{\mathbf{k}\lambda} \hbar/(2\epsilon_0\omega_{\mathbf{k}}V) \neq 0$. This is the same complementarity as the coherent state in reverse: the vacuum concentrates the number-operator Born distribution ($n = 0$ with certainty) and spreads the field-operator Born distribution (zero-point variance).

The nonzero variance is what “vacuum fluctuations” refers to: it is a property of the Born distribution for the field observable in the state $|0\rangle$, not a temporal process in the Fock layer. The vacuum does not fluctuate. It is one quiet vector. The Born distribution for one particular observable happens to have nonzero width.

When interactions are present (QED, the Standard Model), the interacting vacuum $|\Omega\rangle$ is the ground state of the full H : a single vector, expanded in the free Fock basis with components in multi-particle sectors. These components are basis-decomposition coefficients, not Born-layer events—the same point made for the harmonic oscillator in section 6.5.

14.11 The RNN picture of the quantized EM field

The quantized EM field slots into the Fock-Born architecture:

Free photons: diagonal rotation. The free Hamiltonian is diagonal in the Fock basis. Each basis vector $|\{n_{\mathbf{k}\lambda}\}\rangle$ rotates at frequency $\sum n_{\mathbf{k}\lambda}\omega_{\mathbf{k}}$. No amplitude flows between photon-number configurations: independent oscillators, each rotating at its own rate.

Interaction with matter: off-diagonal coupling. The interaction Hamiltonian $H_{\text{int}} = -e\hat{\mathbf{A}} \cdot \hat{\mathbf{p}}/mc$ contains $a_{\mathbf{k}\lambda}a_{\text{electron}}^\dagger$ (photon absorbed, electron excited) and $a_{\mathbf{k}\lambda}^\dagger a_{\text{electron}}$ (photon emitted, electron drops). These transfer amplitude between Fock sectors: the RNN becomes non-trivial.

Classical EM	Quantum EM (Fock-Born)
Vector potential $\mathbf{A}(\mathbf{x}, t)$	Operator $\hat{\mathbf{A}}(\mathbf{x})$ on Fock space
Mode amplitude $\alpha_{\mathbf{k}\lambda} \in \mathbb{C}$	Operator $a_{\mathbf{k}\lambda}$
$ \alpha ^2 = \text{energy in mode}$	$\langle \hat{n} \rangle = \text{mean photon number}$
Classical wave	Coherent state $ \alpha\rangle$
Energy continuous	Quantised: $n\hbar\omega$
Maxwell equations	$\dot{h}_t = -iH h_t$ on $\mathcal{F}_{\text{photon}}$

14.12 Dirac's original construction: from action-angle variables to creation operators

The creation and annihilation operators originate in Dirac's 1927 paper. Understanding his route illuminates what the operators *are*.

Classical action-angle variables. Each mode of the EM field is a harmonic oscillator with action J (proportional to energy: $E = \omega J$) and angle $\theta \in [0, 2\pi)$ (the oscillation phase), canonically conjugate: $\{J, \theta\}_{\text{PB}} = 1$. The complex amplitude is $\alpha = \sqrt{J} e^{i\theta}$.

Quantization. Promote to operators: $[\hat{\theta}, \hat{N}] = i$ with $\hat{N} = \hat{J}/\hbar$. Define $e^{\pm i\hat{\theta}}$; from the commutation relation: $\hat{N} e^{i\hat{\theta}} = e^{i\hat{\theta}}(\hat{N} - 1)$ and $\hat{N} e^{-i\hat{\theta}} = e^{-i\hat{\theta}}(\hat{N} + 1)$. These are ladder operators stepping $\hat{N}$ by ± 1 .

Since $\hat{N} \geq 0$ (the action is non-negative), the lowering must terminate at $|0\rangle$ with $\hat{N}|0\rangle = 0$. Applying $e^{-i\hat{\theta}}$ repeatedly generates $|1\rangle, |2\rangle, \dots$. The eigenvalues of $\hat{N}$ are $n = 0, 1, 2, \dots$: non-negative integers. Energy: $E_n = n\hbar\omega$. Dirac derived Planck's quantization from the algebra of q-numbers.

Creation and annihilation operators. The properly normalised versions: $a = e^{i\hat{\theta}}\sqrt{\hat{N}}$, $a^\dagger = \sqrt{\hat{N} + 1}e^{-i\hat{\theta}}$, satisfying $a|n\rangle = \sqrt{n}|n-1\rangle$, $a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle$, and $[a, a^\dagger] = 1$. The classical amplitude $\alpha = \sqrt{J} e^{i\theta}$ has become the quantum operator $a = e^{i\hat{\theta}}\sqrt{\hat{N}}$: the annihilation operator is the quantized mode amplitude.

What Dirac achieved. In one paper: quantized the radiation field, derived photons from algebra (not postulate), invented creation and annihilation operators, derived Einstein's A and B coefficients for spontaneous and stimulated emission from first principles, and founded quantum field theory. The route from Dirac's action-angle variables to the modern Fock space is a direct line: $\alpha = \sqrt{J} e^{i\theta} \rightarrow a = e^{i\hat{\theta}}\sqrt{\hat{N}} \rightarrow a_{\mathbf{k}\lambda}$ with $[a, a^\dagger] = 1$. From there, the entire technology of QFT follows.

15 Many-Particle Quantum Mechanics: Fixed Particle Number

The Dirac equation introduced us to the single-particle Hilbert space $\mathcal{H}^{(1)} = L^2(\mathbb{R}^3) \otimes \mathbb{C}^4$: basis vectors $|x, A\rangle$ labelled by position and spinor index, with the hidden state a superposition over these. A real physical system — a helium atom with two electrons, a proton with three quarks, a gas of 10^{23} molecules — contains multiple particles. This section develops the quantum mechanics of exactly n identical particles: the n -particle Hilbert space $\mathcal{H}^{(n)}$, its basis vectors, the wavefunction as a projection of h_t onto these basis vectors, and the fundamental distinction between fermions and bosons. This is the halfway house between one-particle quantum mechanics and the full Fock space: particle number is fixed at n , but $n > 1$.

15.1 The naive n -particle space and its problem

For n *distinguishable* particles, the joint Hilbert space is the tensor product:

$$\mathcal{H}_{\text{naive}}^{(n)} = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_n, \quad (15.1)$$

with basis vectors:

$$|x_1, s_1\rangle \otimes |x_2, s_2\rangle \otimes \cdots \otimes |x_n, s_n\rangle \equiv |x_1, s_1; x_2, s_2; \dots; x_n, s_n\rangle, \quad (15.2)$$

where particle k is at position x_k with spin (or spinor index) s_k . The hidden state is a superposition:

$$h_t = \int \prod_{k=1}^n d^3x_k \sum_{s_1, \dots, s_n} \Psi(x_1, s_1; \dots; x_n, s_n; t) |x_1, s_1; \dots; x_n, s_n\rangle, \quad (15.3)$$

and the n -particle wavefunction:

$$\Psi(x_1, s_1; \dots; x_n, s_n; t) = \langle x_1, s_1; \dots; x_n, s_n | h_t \rangle \quad (15.4)$$

is the projection of h_t onto the basis vector labelled by the n -particle configuration. The normalization condition is:

$$\int \prod_k d^3x_k \sum_{s_1, \dots, s_n} |\Psi(x_1, s_1; \dots; x_n, s_n; t)|^2 = 1. \quad (15.5)$$

The problem: particles are identical. In classical mechanics, identical particles can be distinguished by tracking their trajectories: even if two particles have identical properties, we can label them 1 and 2 and follow each one. In quantum mechanics, there are no trajectories: the wavefunction $\Psi(x_1, s_1; x_2, s_2)$ is a function of two position-spin labels, but there is no physical meaning to the statement “particle 1 is at x_1 and particle 2 is at x_2 ” for identical particles. The labels 1 and 2 are a bookkeeping device, not a physical distinction.

This means the physical states cannot depend on which label we assigned to which particle. Exchanging labels 1 and 2 — swapping $(x_1, s_1) \leftrightarrow (x_2, s_2)$ — must map the physical state to itself, up to a phase. Applying the exchange twice returns the original labelling, so the phase $e^{i\theta}$ must satisfy $e^{2i\theta} = 1$, giving $e^{i\theta} = \pm 1$. This is the fundamental dichotomy of quantum statistics.

15.2 Bosons and fermions: the exchange symmetry

For n identical particles, the physical states fall into two classes determined by the behaviour of the wavefunction under the exchange of any two particle labels:

Bosons: the wavefunction is *symmetric* under any exchange:

$$\Psi(\dots, x_i, s_i; \dots; x_j, s_j; \dots) = +\Psi(\dots, x_j, s_j; \dots; x_i, s_i; \dots). \quad (15.6)$$

Bosons have integer spin $(0, 1, 2, \dots)$. Examples: photons ($s = 1$), the Higgs ($s = 0$), the graviton ($s = 2$), ${}^4\text{He}$ nuclei ($s = 0$).

Fermions: the wavefunction is *antisymmetric* under any exchange:

$$\Psi(\dots, x_i, s_i; \dots; x_j, s_j; \dots) = -\Psi(\dots, x_j, s_j; \dots; x_i, s_i; \dots). \quad (15.7)$$

Fermions have half-integer spin $(\frac{1}{2}, \frac{3}{2}, \dots)$. Examples: electrons, quarks, neutrinos, protons (composite, spin $\frac{1}{2}$).

The spin-statistics theorem guarantees that the exchange symmetry is exactly correlated with the spin: integer spin particles are bosons, half-integer spin particles are fermions. This is not an additional postulate; it follows from Lorentz invariance and the positivity of energy.

15.3 The physical n -particle Hilbert spaces

The physical Hilbert space for n identical bosons or fermions is the appropriate symmetry-projected subspace of the naive space (15.1).

Symmetric subspace for bosons. Define the *symmetrisation operator*:

$$\mathcal{S} = \frac{1}{n!} \sum_{\sigma \in S_n} \hat{\sigma}, \quad (15.8)$$

where S_n is the group of permutations of n elements and $\hat{\sigma}$ permutes the particle labels. The bosonic n -particle Hilbert space is:

$$\mathcal{H}_{\text{boson}}^{(n)} = \mathcal{S} \mathcal{H}_{\text{naive}}^{(n)} \equiv \mathcal{H}_1 \vee \mathcal{H}_2 \vee \dots \vee \mathcal{H}_n, \quad (15.9)$$

the symmetric tensor product (also written $\text{Sym}^n \mathcal{H}$).

Antisymmetric subspace for fermions. Define the *antisymmetrisation operator*:

$$\mathcal{A} = \frac{1}{n!} \sum_{\sigma \in S_n} \text{sgn}(\sigma) \hat{\sigma}, \quad (15.10)$$

where $\text{sgn}(\sigma) = +1$ for even permutations and -1 for odd permutations. The fermionic n -particle Hilbert space is:

$$\mathcal{H}_{\text{fermion}}^{(n)} = \mathcal{A} \mathcal{H}_{\text{naive}}^{(n)} \equiv \mathcal{H}_1 \wedge \mathcal{H}_2 \wedge \dots \wedge \mathcal{H}_n, \quad (15.11)$$

the antisymmetric tensor product (also written $\Lambda^n \mathcal{H}$ or the exterior product).

15.4 Basis vectors for n identical particles

Single-particle basis. Let $\{|\lambda\rangle\}$ be an orthonormal basis of the single-particle Hilbert space $\mathcal{H}$, where λ labels all single-particle quantum numbers (momentum $\mathbf{k}$, spin s , colour α , etc.). We work in the momentum-spin basis: $\lambda = (\mathbf{k}, s)$.

Bosonic basis vectors: symmetrised products. A bosonic n -particle basis vector is a symmetrised product of single-particle states:

$$|\lambda_1, \lambda_2, \dots, \lambda_n\rangle_{\text{boson}} = \sqrt{\frac{n!}{\prod_{\lambda} n_{\lambda}!}} \mathcal{S} (|\lambda_1\rangle \otimes |\lambda_2\rangle \otimes \dots \otimes |\lambda_n\rangle), \quad (15.12)$$

where n_λ is the number of particles in state λ (the occupation number) and the prefactor ensures normalization. The basis vectors are labelled by the *multiset* $\{\lambda_1, \dots, \lambda_n\}$: the set of single-particle states occupied, with repetition allowed (multiple bosons can occupy the same state).

Concrete two-boson example. For two bosons with $\lambda_1 = \mathbf{k}_1$ and $\lambda_2 = \mathbf{k}_2$:

$$|\mathbf{k}_1, \mathbf{k}_2\rangle_{\text{boson}} = \frac{1}{\sqrt{2}}(|\mathbf{k}_1\rangle \otimes |\mathbf{k}_2\rangle + |\mathbf{k}_2\rangle \otimes |\mathbf{k}_1\rangle) \quad (\mathbf{k}_1 \neq \mathbf{k}_2), \quad (15.13)$$

$$|\mathbf{k}, \mathbf{k}\rangle_{\text{boson}} = |\mathbf{k}\rangle \otimes |\mathbf{k}\rangle \quad (\text{both bosons in the same state: allowed}). \quad (15.14)$$

Two bosons can occupy the same momentum state: the symmetric wavefunction is non-zero when $\mathbf{k}_1 = \mathbf{k}_2$.

Fermionic basis vectors: Slater determinants. A fermionic n -particle basis vector is an antisymmetrised product of single-particle states:

$$|\lambda_1, \lambda_2, \dots, \lambda_n\rangle_{\text{fermion}} = \frac{1}{\sqrt{n!}} \sum_{\sigma \in S_n} \text{sgn}(\sigma) |\lambda_{\sigma(1)}\rangle \otimes |\lambda_{\sigma(2)}\rangle \otimes \dots \otimes |\lambda_{\sigma(n)}\rangle. \quad (15.15)$$

This is the *Slater determinant*: the wavefunction in position space is:

$$\Psi(x_1, s_1; \dots; x_n, s_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_{\lambda_1}(x_1, s_1) & \phi_{\lambda_1}(x_2, s_2) & \dots & \phi_{\lambda_1}(x_n, s_n) \\ \phi_{\lambda_2}(x_1, s_1) & \phi_{\lambda_2}(x_2, s_2) & \dots & \phi_{\lambda_2}(x_n, s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{\lambda_n}(x_1, s_1) & \phi_{\lambda_n}(x_2, s_2) & \dots & \phi_{\lambda_n}(x_n, s_n) \end{vmatrix}, \quad (15.16)$$

where $\phi_\lambda(x, s) = \langle x, s | \lambda \rangle$ is the single-particle wavefunction for state λ .

Concrete two-fermion example. For two fermions with $\lambda_1 = (\mathbf{k}_1, s_1)$ and $\lambda_2 = (\mathbf{k}_2, s_2)$:

$$|(\mathbf{k}_1, s_1); (\mathbf{k}_2, s_2)\rangle_{\text{fermion}} = \frac{1}{\sqrt{2}}(|\mathbf{k}_1, s_1\rangle \otimes |\mathbf{k}_2, s_2\rangle - |\mathbf{k}_2, s_2\rangle \otimes |\mathbf{k}_1, s_1\rangle). \quad (15.17)$$

If $\lambda_1 = \lambda_2$ (both fermions in the same state):

$$|\lambda, \lambda\rangle_{\text{fermion}} = \frac{1}{\sqrt{2}}(|\lambda\rangle \otimes |\lambda\rangle - |\lambda\rangle \otimes |\lambda\rangle) = 0. \quad (15.18)$$

The antisymmetric basis vector vanishes identically when two particles share the same quantum numbers. This is the *Pauli exclusion principle*: no two fermions can occupy the same single-particle state.

15.5 The Pauli exclusion principle

The Pauli exclusion principle is not an independent postulate. It is a *theorem*: a direct consequence of the antisymmetry requirement (15.7).

Proof. Suppose two fermions are in the same single-particle state λ : $\lambda_i = \lambda_j = \lambda$ for some $i \neq j$. Exchanging particle labels i and j leaves the physical configuration unchanged (the particles are identical and in the same state), so the wavefunction must equal itself:

$$\Psi(\dots, \lambda_i, \dots, \lambda_j, \dots) = \Psi(\dots, \lambda_j, \dots, \lambda_i, \dots) = \Psi(\dots, \lambda, \dots, \lambda, \dots). \quad (15.19)$$

But antisymmetry requires:

$$\Psi(\dots, \lambda_i, \dots, \lambda_j, \dots) = -\Psi(\dots, \lambda_j, \dots, \lambda_i, \dots). \quad (15.20)$$

Together: $\Psi = -\Psi$, so $\Psi = 0$. The wavefunction vanishes: the state does not exist. No two fermions can occupy the same single-particle state. $\square$

Physical consequences. The Pauli exclusion principle is responsible for the structure of matter. Without it:

- All electrons in an atom would collapse to the lowest energy level $1s$, and there would be no periodic table.
- Neutron stars would collapse: the *neutron degeneracy pressure* (the resistance of neutrons to being forced into the same quantum state) is what supports neutron stars against gravitational collapse.
- Ordinary matter would not be stable: atoms would collapse to nuclear densities.

The stability of matter is a consequence of the antisymmetry of the fermionic wavefunction — a property of the Fock basis vectors, not of any interaction between particles.

Bosonic enhancement. Bosons have the opposite tendency: many bosons in the same state is not forbidden but *enhanced*. The probability amplitude for adding a boson to a state already containing n bosons is proportional to $\sqrt{n+1}$ (from the creation operator, as we will see in the Fock space formalism). This enhancement drives *Bose-Einstein condensation*: below a critical temperature, a macroscopic fraction of bosons condenses into the same single-particle ground state, forming a quantum collective mode. Lasers (photon BEC), superfluids (^4He), and superconductors (Cooper pair BEC) all exploit bosonic enhancement.

15.6 The n -particle wavefunction as a projection

The hidden state $h_t \in \mathcal{H}^{(n)}$ (restricted to the n -particle sector) contains the full quantum information about n identical particles. The wavefunction is the projection of h_t onto the n -particle basis vectors.

For fermions. Inserting the resolution of the identity in the fermionic basis:

$$h_t = \frac{1}{n!} \sum_{s_1, \dots, s_n} \int \prod_{k=1}^n d^3x_k \Psi(x_1, s_1; \dots; x_n, s_n; t) |x_1, s_1; \dots; x_n, s_n\rangle_{\text{fermion}}, \quad (15.21)$$

where the antisymmetric wavefunction satisfies:

$$\Psi(\dots; x_i, s_i; \dots; x_j, s_j; \dots) = -\Psi(\dots; x_j, s_j; \dots; x_i, s_i; \dots). \quad (15.22)$$

The factor of $1/n!$ corrects for the overcounting from the sum over all orderings. The probability density for finding particle 1 at x_1 with spin s_1 , and particle 2 at x_2 with spin s_2 , etc. is:

$$P(x_1, s_1; \dots; x_n, s_n) = |\Psi(x_1, s_1; \dots; x_n, s_n; t)|^2. \quad (15.23)$$

For bosons. The bosonic wavefunction is symmetric and the expansion is analogous, with $+$ signs throughout. The occupation number representation — specifying how many particles are in each single-particle state $\{\lambda\}$ — is often more convenient:

$$h_t = \sum_{\{n_\lambda\}: \sum_\lambda n_\lambda = n} c(\{n_\lambda\}; t) |\{n_\lambda\}\rangle_{\text{boson}}, \quad (15.24)$$

where $\{n_\lambda\}$ specifies the occupation number of each single-particle state, and $c(\{n_\lambda\}; t) = \langle \{n_\lambda\} | h_t \rangle$ is the Born amplitude for the configuration.

Born rendering for n -particle states. The Born rendering layer for the position observable $\hat{X}_1 \otimes \cdots \otimes \hat{X}_n$ gives the probability density:

$$P(x_1, \dots, x_n) = \sum_{s_1, \dots, s_n} |\Psi(x_1, s_1; \dots; x_n, s_n; t)|^2, \quad (15.25)$$

the probability of finding the n particles at positions $x_1, \dots, x_n$ (summed over spins). For fermions, the antisymmetry of Ψ ensures that this probability vanishes when $x_i = x_j$ (for particles with the same spin), reflecting the fact that two identical fermions cannot be at the same place with the same quantum numbers.

15.7 The Hamiltonian motor for n identical particles

The Schrödinger equation $\dot{h}_t = -iH^{(n)}h_t$ governs the evolution of the n -particle hidden state. The n -particle Hamiltonian has two types of terms:

One-body terms. A term that acts on each particle independently:

$$H_{1\text{-body}}^{(n)} = \sum_{k=1}^n \hat{h}_k, \quad \hat{h}_k = -\frac{\nabla_k^2}{2m} + V(x_k), \quad (15.26)$$

where $\hat{h}_k$ acts on the k -th particle's position and spin. For a symmetric (bosonic) or antisymmetric (fermionic) wavefunction, this acts consistently: $H_{1\text{-body}}^{(n)}$ commutes with the symmetrisation and antisymmetrisation operators.

Two-body interaction terms. A term that involves pairs of particles:

$$H_{2\text{-body}}^{(n)} = \frac{1}{2} \sum_{k \neq l} V_{\text{int}}(x_k - x_l), \quad (15.27)$$

where V_{int} is the interaction potential (e.g. the Coulomb repulsion $e^2/|x_k - x_l|$ between electrons). The factor of $\frac{1}{2}$ prevents double counting. The two-body term couples different single-particle states: it is the source of correlations in the n -particle wavefunction.

15.7.1 Eigenstructure of the one-body Hamiltonian

The single-particle Hamiltonian $\hat{h} = -\nabla^2/2m + V(x)$ has eigenstates $\{\phi_\alpha\}$ and eigenvalues $\{\varepsilon_\alpha\}$:

$$\hat{h} \phi_\alpha(x) = \varepsilon_\alpha \phi_\alpha(x), \quad \alpha = 1, 2, 3, \dots \quad (15.28)$$

The label α is a composite index: for the hydrogen-like atom, $\alpha = (n, l, m_l, m_s)$, where n is the principal quantum number, l the orbital angular momentum, m_l the magnetic quantum number, and m_s the spin projection. The eigenstates form a complete orthonormal basis of the single-particle Hilbert space: $\langle \phi_\alpha | \phi_\beta \rangle = \delta_{\alpha\beta}$, $\sum_\alpha |\phi_\alpha\rangle \langle \phi_\alpha| = \mathbf{1}$.

For a Coulomb potential $V(x) = -Ze^2/|x|$ (the hydrogen-like atom), the eigenvalues are:

$$\varepsilon_{nlm_l m_s} = -\frac{Z^2 m e^4}{2\hbar^2 n^2}, \quad n = 1, 2, 3, \dots \quad (15.29)$$

Each energy level n has degeneracy $2n^2$ (the factor of 2 from spin, n^2 from the angular momentum quantum numbers $l = 0, \dots, n-1$ and $m_l = -l, \dots, l$).

These single-particle eigenstates are the **orbitals**: the building blocks from which n -particle states are constructed.

15.7.2 Eigenstructure of the non-interacting n -particle system

When the interaction is absent ($H_{2\text{-body}}^{(n)} = 0$), the n -particle Hamiltonian is purely one-body: $H_0^{(n)} = \sum_{k=1}^n \hat{h}_k$. The eigenvalue problem separates: the n -particle eigenstates are products of single-particle eigenstates, and the n -particle eigenvalues are sums of single-particle eigenvalues.

Distinguishable particles. If the particles were distinguishable, the eigenstates would be simple products:

$$\Phi_{\alpha_1 \dots \alpha_n}(x_1, \dots, x_n) = \phi_{\alpha_1}(x_1) \phi_{\alpha_2}(x_2) \cdots \phi_{\alpha_n}(x_n), \quad (15.30)$$

with eigenvalue $E = \varepsilon_{\alpha_1} + \varepsilon_{\alpha_2} + \cdots + \varepsilon_{\alpha_n}$. But identical particles are not distinguishable.

Fermions: Slater determinants. For n identical fermions, the wavefunction must be antisymmetric under the exchange of any two particles. The antisymmetrised product of n single-particle orbitals is the **Slater determinant**:

$$\Psi_{\alpha_1 \dots \alpha_n}(x_1, \dots, x_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_{\alpha_1}(x_1) & \phi_{\alpha_1}(x_2) & \cdots & \phi_{\alpha_1}(x_n) \\ \phi_{\alpha_2}(x_1) & \phi_{\alpha_2}(x_2) & \cdots & \phi_{\alpha_2}(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{\alpha_n}(x_1) & \phi_{\alpha_n}(x_2) & \cdots & \phi_{\alpha_n}(x_n) \end{vmatrix}. \quad (15.31)$$

The determinant automatically enforces antisymmetry: swapping two particles (two columns) changes the sign; putting two particles in the same orbital (two identical rows) gives zero. The Pauli exclusion principle is built into the determinant.

The eigenvalue of the Slater determinant under $H_0^{(n)}$ is the sum of the occupied orbital energies:

$$E_{\alpha_1 \dots \alpha_n}^{(0)} = \varepsilon_{\alpha_1} + \varepsilon_{\alpha_2} + \cdots + \varepsilon_{\alpha_n}. \quad (15.32)$$

The non-interacting ground state is the Slater determinant built from the n lowest-energy orbitals: the electrons fill the orbitals from the bottom up, two per spatial orbital (one spin-up, one spin-down), producing the shell structure of atoms.

Bosons: permanents. For n identical bosons, the wavefunction must be symmetric under exchange. The symmetrised product is the **permanent**:

$$\Psi_{\alpha_1 \dots \alpha_n}^{(\text{sym})}(x_1, \dots, x_n) = \frac{1}{\sqrt{n! \prod_{\alpha} n_{\alpha}!}} \sum_{\sigma \in S_n} \phi_{\alpha_{\sigma(1)}}(x_1) \cdots \phi_{\alpha_{\sigma(n)}}(x_n), \quad (15.33)$$

where the sum is over all permutations σ of $\{1, \dots, n\}$, and n_{α} is the number of particles in orbital α . Unlike the determinant, the permanent does not vanish when two particles occupy the same orbital: bosons can pile into the same state. The non-interacting ground state of n bosons has all n particles in the lowest orbital: $\Psi_{\text{gs}} = \phi_1(x_1)\phi_1(x_2)\cdots\phi_1(x_n)$ —the Bose-Einstein condensate in its simplest form.

The non-interacting eigenstructure, summarised.

	Fermions	Bosons
n -particle eigenstate	Slater determinant of n orbitals	Permanent of n orbitals
Eigenvalue	$\sum_{k=1}^n \varepsilon_{\alpha_k}$	$\sum_{k=1}^n \varepsilon_{\alpha_k}$
Occupation per orbital	0 or 1 (Pauli exclusion)	0, 1, 2, ... (no restriction)
Ground state	Fill lowest n orbitals	All n in lowest orbital
Structure	Shell filling, periodic table	Condensation

The Slater determinants (fermions) or permanents (bosons) are the eigenstates of the non-interacting Hamiltonian $H_0^{(n)}$. They form a complete basis for the n -particle Hilbert space (restricted to the antisymmetric or symmetric subspace, respectively). Any n -particle state can be expanded in this basis.

15.7.3 How interactions change the eigenstructure

The two-body interaction $H_{2\text{-body}}^{(n)}$ is off-diagonal in the Slater determinant (or permanent) basis: it couples different orbital configurations, transferring particles between orbitals. The eigenstates of the full Hamiltonian $H^{(n)} = H_0^{(n)} + H_{2\text{-body}}^{(n)}$ are *not* Slater determinants. They are superpositions of Slater determinants—correlated states in which the particles' positions are entangled by the interaction.

Configuration interaction. The exact n -particle eigenstate can be expanded in the basis of Slater determinants:

$$\Psi_{\text{exact}} = \sum_{\{\alpha_1 < \dots < \alpha_n\}} C_{\alpha_1 \dots \alpha_n} \Psi_{\alpha_1 \dots \alpha_n}, \quad (15.34)$$

where the sum is over all ways of choosing n orbitals from the complete set, and $C_{\alpha_1 \dots \alpha_n}$ are the expansion coefficients. This expansion is called **configuration interaction** (CI): the exact state is a superposition of configurations (Slater determinants), each representing a specific assignment of particles to orbitals.

The CI expansion is exact if the sum runs over all configurations (full CI). In practice, the number of configurations grows combinatorially: for n electrons in M orbitals, the number of Slater determinants is $\binom{M}{n}$, which for $M = 100$ and $n = 10$ is $\sim 10^{13}$. This exponential growth is the **many-body problem**: the exact solution of the interacting n -particle system is exponentially expensive in the number of particles and orbitals.

This is the exponential abyss in the language of many-body physics. The hidden state h_t lives in a Hilbert space of dimension $\binom{M}{n}$ (for fermions), and the classical description of this state requires $\binom{M}{n}$ complex numbers. The Born rendering—the eigenvalue of an observable—is a single number projected from this exponentially large space. The gap between h_t and the rendered output is the gap between $\binom{M}{n}$ dimensions and one number.

What the interaction does, physically. The non-interacting ground state of the helium atom (for example) has both electrons in the $1s$ orbital: a single Slater determinant. The interaction $e^2/|x_1 - x_2|$ mixes this configuration with excited configurations— $1s\,2s$, $1s\,2p$, $1s\,3d$, $2s\,2p$, $\dots$ —producing a correlated ground state:

$$\Psi_{\text{gs}}^{\text{He}} = C_0 \Psi_{1s,1s} + C_1 \Psi_{1s,2s} + C_2 \Psi_{1s,2p} + C_3 \Psi_{2s,2s} + \dots \quad (15.35)$$

The dominant coefficient C_0 is close to 1 (the non-interacting configuration is a good approximation), but the corrections $C_1, C_2, C_3, \dots$ are nonzero and physically important: they account for the **correlation energy**—the energy lowering due to the electrons avoiding each other (“keeping out of each other’s way”), which the single-determinant approximation cannot capture.

The interaction creates *entanglement* between the particles: in the correlated ground state (15.35), the position of electron 1 is correlated with the position of electron 2. If electron 1 is found on the left, electron 2 is more likely to be on the right, because the Coulomb repulsion penalises configurations where both electrons are close together. This correlation is encoded in the off-diagonal coefficients $C_1, C_2, \dots$, which mix different Slater determinants.

15.7.4 The helium atom: a complete example

The helium atom ($n = 2$, $Z = 2$) is the simplest system that exhibits the full complexity of the interacting eigenstructure. The Hamiltonian:

$$H_{\text{He}}^{(2)} = \underbrace{-\frac{\nabla_1^2}{2m} - \frac{Ze^2}{|x_1|}}_{\hat{h}_1} + \underbrace{-\frac{\nabla_2^2}{2m} - \frac{Ze^2}{|x_2|}}_{\hat{h}_2} + \underbrace{\frac{e^2}{|x_1 - x_2|}}_{\text{electron-electron repulsion}}. \quad (15.36)$$

The non-interacting eigenstructure (e - e repulsion = 0). Without the electron-electron interaction, the eigenstates are products of hydrogenic orbitals with $Z = 2$. The ground state energy is:

$$E_{\text{gs}}^{(0)} = 2 \times \left(-\frac{Z^2 m e^4}{2\hbar^2} \right) = -4 \text{ Ry} = -108.8 \text{ eV}, \quad (15.37)$$

where 1 Rydberg = 13.6 eV. The experimental ground state energy is -79.0 eV: the non-interacting estimate is off by 38%. The electron-electron repulsion raises the energy substantially.

First-order perturbation theory. Treating $e^2/|x_1 - x_2|$ as a perturbation, the first-order energy correction is:

$$\delta E^{(1)} = \langle \Psi_{\text{gs}}^{(0)} | \frac{e^2}{|x_1 - x_2|} | \Psi_{\text{gs}}^{(0)} \rangle = \frac{5}{4} Z \text{ Ry} = \frac{5}{8} \times 4 \text{ Ry} = +34.0 \text{ eV}. \quad (15.38)$$

The corrected ground state energy is $E^{(0)} + \delta E^{(1)} = -108.8 + 34.0 = -74.8$ eV, within 5% of the experimental value -79.0 eV. Higher-order perturbation theory and variational methods bring the theoretical value into agreement with experiment to arbitrary precision.

The ground state wavefunction: antisymmetry and spin. The two-electron wavefunction must be antisymmetric under exchange $1 \leftrightarrow 2$. It factorises into a spatial part and a spin part. For the ground state, both electrons occupy the $1s$ orbital (symmetric spatial part), so the spin part must be antisymmetric:

$$\Psi_{\text{gs}}(x_1, s_1; x_2, s_2) = \underbrace{\phi_{1s}(x_1)\phi_{1s}(x_2)}_{\text{symmetric in space}} \cdot \underbrace{\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)}_{\text{antisymmetric in spin: singlet}}. \quad (15.39)$$

The spin state is the **singlet**: total spin $S = 0$, $m_S = 0$. The two electrons have opposite spins. This is the only way to put two fermions in the same spatial orbital: the Pauli exclusion principle is satisfied because the composite quantum numbers $(1s, \uparrow)$ and $(1s, \downarrow)$ are distinct.

Excited states: singlet and triplet. When one electron is promoted to a higher orbital (say $2s$ or $2p$), the two electrons occupy different spatial orbitals, and two spin configurations are possible:

Singlet ($S = 0$): antisymmetric spin, symmetric spatial.

$$\Psi_{\text{singlet}} = \frac{1}{\sqrt{2}}[\phi_a(x_1)\phi_b(x_2) + \phi_a(x_2)\phi_b(x_1)] \cdot \frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle). \quad (15.40)$$

Triplet ($S = 1$): symmetric spin, antisymmetric spatial.

$$\Psi_{\text{triplet}} = \frac{1}{\sqrt{2}}[\phi_a(x_1)\phi_b(x_2) - \phi_a(x_2)\phi_b(x_1)] \cdot \chi_{\text{triplet}}(s_1, s_2), \quad (15.41)$$

where χ_{triplet} is one of the three symmetric spin states: $|\uparrow\uparrow\rangle$, $\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle)$, or $|\downarrow\downarrow\rangle$.

The exchange splitting. The singlet and triplet states of the same orbital configuration (a, b) have *different energies*, even though they involve the same orbitals. The energy difference comes from the electron-electron interaction applied to the two different spatial wavefunctions. Define:

$$J_{ab} = \int \int \frac{|\phi_a(x_1)|^2 |\phi_b(x_2)|^2}{|x_1 - x_2|} d^3x_1 d^3x_2 \quad (\text{Coulomb integral}), \quad (15.42)$$

$$K_{ab} = \int \int \frac{\phi_a^*(x_1) \phi_b^*(x_2) \phi_b(x_1) \phi_a(x_2)}{|x_1 - x_2|} d^3x_1 d^3x_2 \quad (\text{exchange integral}). \quad (15.43)$$

The Coulomb integral J_{ab} is the classical electrostatic repulsion between the charge distributions $|\phi_a|^2$ and $|\phi_b|^2$. The exchange integral K_{ab} has no classical analogue: it arises from the overlap of the two orbitals in the interaction region and is a purely quantum-mechanical effect.

The energies are:

$$E_{\text{singlet}} = \varepsilon_a + \varepsilon_b + J_{ab} + K_{ab}, \quad (15.44)$$

$$E_{\text{triplet}} = \varepsilon_a + \varepsilon_b + J_{ab} - K_{ab}. \quad (15.45)$$

The exchange integral $K_{ab} > 0$ (for the Coulomb interaction), so:

$$E_{\text{triplet}} < E_{\text{singlet}}. \quad (15.46)$$

The triplet lies below the singlet. This is the **exchange splitting**: a purely quantum effect, arising from the antisymmetry of the wavefunction, with no classical counterpart.

The physical reason: in the triplet state, the spatial wavefunction is antisymmetric—it vanishes when $x_1 = x_2$. The two electrons avoid each other. The Coulomb repulsion is reduced. In the singlet state, the spatial wavefunction is symmetric—it does not vanish when $x_1 = x_2$. The electrons are more likely to be found near each other. The Coulomb repulsion is larger.

The exchange splitting is the origin of Hund's first rule in atomic physics (maximise total spin to minimise energy), and it is the mechanism behind ferromagnetism: in a solid, the exchange interaction between neighbouring atoms favours parallel spins (triplet), aligning the magnetic moments and producing a permanent magnet.

The helium spectrum, summarised.

Configuration	Spin	Energy	Degeneracy
$1s^2$ (ground)	Singlet ($S = 0$)	$2\varepsilon_{1s} + J_{1s,1s}$	1
$1s2s$	Triplet ($S = 1$)	$\varepsilon_{1s} + \varepsilon_{2s} + J - K$	3
$1s2s$	Singlet ($S = 0$)	$\varepsilon_{1s} + \varepsilon_{2s} + J + K$	1
$1s2p$	Triplet ($S = 1$)	$\varepsilon_{1s} + \varepsilon_{2p} + J' - K'$	9
$1s2p$	Singlet ($S = 0$)	$\varepsilon_{1s} + \varepsilon_{2p} + J' + K'$	3

The helium spectrum splits into two families: **parahelium** (singlet, $S = 0$) and **orthohelium** (triplet, $S = 1$), with the triplet states systematically below the corresponding singlet states. The ground state ($1s^2$) is necessarily a singlet (both electrons in the same orbital forces antisymmetric spin). The first excited states ($1s2s$) split into a triplet at -59.2 eV and a singlet at -58.4 eV: the exchange splitting is ~ 0.8 eV, a measurable and well-confirmed effect.

15.7.5 From n -particle wavefunctions to Fock space

The eigenstructure of the n -particle Hamiltonian—Slater determinants for the non-interacting theory, configuration interaction for the interacting theory—is the first-quantized language. The

second-quantized language reformulates the same physics in terms of creation and annihilation operators on the Fock space.

The correspondence is direct:

- The single-particle orbital ϕ_α corresponds to the creation operator $c_\alpha^\dagger$ that creates a particle in orbital α .
- The Slater determinant $\Psi_{\alpha_1 \dots \alpha_n}$ corresponds to the Fock state $c_{\alpha_1}^\dagger \dots c_{\alpha_n}^\dagger |0\rangle$.
- The one-body Hamiltonian $H_{1\text{-body}} = \sum_k \hat{h}_k$ becomes $\sum_{\alpha\beta} h_{\alpha\beta} c_\alpha^\dagger c_\beta$, where $h_{\alpha\beta} = \langle \phi_\alpha | \hat{h} | \phi_\beta \rangle$.
- The two-body interaction $H_{2\text{-body}}$ becomes $\frac{1}{2} \sum_{\alpha\beta\gamma\delta} V_{\alpha\beta\gamma\delta} c_\alpha^\dagger c_\beta^\dagger c_\delta c_\gamma$, where $V_{\alpha\beta\gamma\delta} = \langle \phi_\alpha \phi_\beta | V_{\text{int}} | \phi_\gamma \phi_\delta \rangle$.

The full n -particle Hamiltonian becomes a polynomial of creation and annihilation operators:

$$H = \sum_{\alpha\beta} h_{\alpha\beta} c_\alpha^\dagger c_\beta + \frac{1}{2} \sum_{\alpha\beta\gamma\delta} V_{\alpha\beta\gamma\delta} c_\alpha^\dagger c_\beta^\dagger c_\delta c_\gamma. \quad (15.47)$$

The first term (degree 2) is the one-body part: it hops particles between orbitals. The second term (degree 4) is the two-body interaction: it scatters pairs of particles between orbitals. This is the Hamiltonian motor for the n -particle system, and it is the same architectural form as the Standard Model motor: a polynomial of creation and annihilation operators, degree ≤ 4 , acting on the Fock space.

The Slater determinant basis of first quantization is the Fock basis of second quantization. Configuration interaction is the expansion of the exact eigenstate in the Fock basis. The exponential growth of the CI expansion is the exponential dimension of the Fock space. The two languages describe the same physics—the same eigenstructure, the same correlations, the same exponential abyss—in different notation.

15.8 Foreshadowing Fock space: connecting particle-number sectors

The n -particle Hilbert spaces $\mathcal{H}^{(n)}$ for different values of n are distinct, orthogonal spaces:

$$\langle m\text{-particle} | n\text{-particle} \rangle = 0 \quad \text{for } m \neq n. \quad (15.48)$$

Each $\mathcal{H}^{(n)}$ describes a world with exactly n particles: basis vectors from different sectors are orthogonal because they differ in the number of particles they describe. Within the n -particle sector, the Schrödinger equation rotates h_t within $\mathcal{H}^{(n)}$, preserving particle number.

Why fixed particle number is insufficient. Physical processes violate fixed particle number. An atom emits a photon: one atom $\rightarrow$ one atom plus one photon. A neutron decays: one neutron $\rightarrow$ one proton plus one electron plus one antineutrino. Pair production: one photon $\rightarrow$ one electron plus one positron. In each case, the particle number changes, and the hidden state must move between sectors. The n -particle Hilbert space $\mathcal{H}^{(n)}$ cannot accommodate this: it is locked to n particles.

The Fock space as a direct sum. The natural resolution is to allow the hidden state to live in the direct sum of all particle-number sectors simultaneously:

$$\mathcal{F} = \mathcal{H}^{(0)} \oplus \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)} \oplus \dots = \bigoplus_{n=0}^{\infty} \mathcal{H}^{(n)}, \quad (15.49)$$

where $\mathcal{H}^{(0)} = \mathbb{C}$ is the vacuum sector (one basis vector: the state with no particles, $|0\rangle$). The hidden state $h_t \in \mathcal{F}$ is a superposition across all sectors:

$$h_t = c_0 |0\rangle \oplus \int d^3k \, c_1(\mathbf{k}) |\mathbf{k}\rangle \oplus \int d^3k_1 d^3k_2 \, c_2(\mathbf{k}_1, \mathbf{k}_2) |\mathbf{k}_1, \mathbf{k}_2\rangle \oplus \dots \quad (15.50)$$

with $|c_0|^2 + \int |c_1|^2 + \int |c_2|^2 + \dots = 1$. Each sector $\mathcal{H}^{(n)}$ is orthogonal to all others (15.48): the n -particle and m -particle pieces of h_t never interfere.

Creation and annihilation operators connect the sectors. The bridge between sectors is provided by the *creation operator* $a^\dagger(\lambda)$ and the *annihilation operator* $a(\lambda)$, which map between adjacent sectors:

$$a^\dagger(\lambda) : \mathcal{H}^{(n)} \rightarrow \mathcal{H}^{(n+1)}, \quad (15.51)$$

$$a(\lambda) : \mathcal{H}^{(n)} \rightarrow \mathcal{H}^{(n-1)}. \quad (15.52)$$

Their action on the basis vectors is:

$$a^\dagger(\lambda) |\lambda_1, \dots, \lambda_n\rangle = \sqrt{n+1} |\lambda, \lambda_1, \dots, \lambda_n\rangle_{\text{sym/antisym}}, \quad (15.53)$$

$$a(\lambda) |\lambda_1, \dots, \lambda_n\rangle = \frac{1}{\sqrt{n}} \sum_{k=1}^n (\pm 1)^{k-1} \delta_{\lambda\lambda_k} |\lambda_1, \dots, \hat{\lambda}_k, \dots, \lambda_n\rangle, \quad (15.54)$$

where $\hat{\lambda}_k$ means λ_k is omitted, and the $\pm$ is $+$ for bosons and $-$ for fermions (from the antisymmetry of the fermionic basis vector). These satisfy:

$$\text{Bosons: } [a(\lambda), a^\dagger(\lambda')] = \delta_{\lambda\lambda'} \mathbf{1}, \quad (15.55)$$

$$\text{Fermions: } \{a(\lambda), a^\dagger(\lambda')\} = \delta_{\lambda\lambda'} \mathbf{1}. \quad (15.56)$$

The bosonic commutator $[a, a^\dagger] = 1$ and the fermionic anticommutator $\{a, a^\dagger\} = 1$ are both consequences of the orthonormality of the Fock basis vectors: they encode, in algebraic form, the normalization and orthogonality of the states $|\lambda_1, \dots, \lambda_n\rangle$. The Pauli exclusion principle appears algebraically as $(a^\dagger(\lambda))^2 = 0$ for fermions (from the anticommutator): applying the creation operator twice for the same state gives zero, because the antisymmetric basis vector vanishes when two particles share the same quantum numbers.

The Hamiltonian in terms of creation and annihilation operators. The one-body and two-body Hamiltonians (15.26) and (15.27) can be written compactly in terms of creation and annihilation operators acting on the full Fock space:

$$H_{1\text{-body}} = \sum_{\lambda, \lambda'} \langle \lambda | \hat{h} | \lambda' \rangle a^\dagger(\lambda) a(\lambda'), \quad (15.57)$$

$$H_{2\text{-body}} = \frac{1}{2} \sum_{\lambda_1, \lambda_2, \lambda'_1, \lambda'_2} \langle \lambda_1, \lambda_2 | V_{\text{int}} | \lambda'_1, \lambda'_2 \rangle a^\dagger(\lambda_1) a^\dagger(\lambda_2) a(\lambda'_2) a(\lambda'_1). \quad (15.58)$$

These are the same physical Hamiltonians (15.26) and (15.27), now written as operators on the full Fock space: they are polynomials in the creation and annihilation operators, and they act correctly on any particle-number sector. When the Hamiltonian includes only (15.57) and (15.58), it preserves particle number: $a^\dagger a$ creates and then destroys, keeping n fixed. Adding terms like $a^\dagger a^\dagger a$ or $a^\dagger$ alone changes the particle number and connects different sectors: these are the interaction vertices of quantum field theory.

Concept	Bosons	Fermions
Exchange symmetry	$\Psi(\dots, i, \dots, j, \dots) + \Psi(\dots, j, \dots, i, \dots)$	$= \Psi(\dots, i, \dots, j, \dots) - \Psi(\dots, j, \dots, i, \dots) =$
Spin	Integer $(0, 1, 2, \dots)$	Half-integer $(\frac{1}{2}, \frac{3}{2}, \dots)$
Basis vectors	Symmetrised products	Slater determinants
Same-state occupation	Allowed (enhanced)	Forbidden (Pauli exclusion)
Algebra	$[a, a^\dagger] = 1$	$\{a, a^\dagger\} = 1, (a^\dagger)^2 = 0$
Physical consequence	BEC, laser, superfluidity	Periodic table, neutron stars, stable matter
Examples	γ , Higgs, ^4He , Cooper pairs	e^- , quarks, ν , protons

The n -particle quantum mechanics developed in this section is the halfway house. The hidden state lives in $\mathcal{H}_{\text{boson}}^{(n)}$ or $\mathcal{H}_{\text{fermion}}^{(n)}$: a fixed sector of the Fock space. The wavefunction is the projection of h_t onto the symmetrised or antisymmetrised basis vectors. The Pauli exclusion principle, Bose-Einstein condensation, and the structure of matter all follow from the symmetry properties of these basis vectors — not from any interaction. The full Fock space $\mathcal{F} = \bigoplus_n \mathcal{H}^{(n)}$ awaits: a direct sum of all these sectors, connected by creation and annihilation operators, with the Hamiltonian of quantum field theory moving amplitude between sectors as particles are created and destroyed.

15.9 Preview: symmetrisation and the origin of quantum statistics

The n -particle Hilbert space constructed in this section used the naive tensor product $\mathcal{H}_{\text{naive}}^{(n)} = \mathcal{H}_1 \otimes \dots \otimes \mathcal{H}_n$, with basis vectors $|x_1, s_1; \dots; x_n, s_n\rangle$ labelled by an ordered list of single-particle states. This labelling assigns a definite identity to each particle: particle 1 is at x_1 , particle 2 is at x_2 , and so on. But identical particles — two electrons, two photons, two up quarks — have no individual identity. There is no physical observable that distinguishes “particle 1” from “particle 2” when the two particles have identical intrinsic properties. The label is a bookkeeping convention, not a physical fact. This observation has profound consequences that we address here, before developing the Fock space formalism that resolves them naturally.

The exchange problem. Swapping the labels of two identical particles — replacing $(x_1, s_1) \leftrightarrow (x_2, s_2)$ — produces a different basis vector in $\mathcal{H}_{\text{naive}}^{(n)}$ but the same physical state. The wavefunction:

$$\Psi(x_1, s_1; \dots; x_i, s_i; \dots; x_j, s_j; \dots) \quad (15.59)$$

and the exchanged wavefunction:

$$\Psi(x_1, s_1; \dots; x_j, s_j; \dots; x_i, s_i; \dots) \quad (15.60)$$

must describe the same physical state. In quantum mechanics, two state vectors describe the same physical state if and only if they differ by a global phase $e^{i\theta}$. Applying the exchange twice returns to the original ordering, so $e^{2i\theta} = 1$, giving $e^{i\theta} = \pm 1$. There are therefore exactly two possibilities, and which one applies is a property of the particle species:

$$\Psi(\dots, x_i, s_i, \dots, x_j, s_j, \dots) = +\Psi(\dots, x_j, s_j, \dots, x_i, s_i, \dots) \quad (\text{bosons}), \quad (15.61)$$

$$\Psi(\dots, x_i, s_i, \dots, x_j, s_j, \dots) = -\Psi(\dots, x_j, s_j, \dots, x_i, s_i, \dots) \quad (\text{fermions}). \quad (15.62)$$

Bosons have symmetric wavefunctions; fermions have antisymmetric wavefunctions. The physical Hilbert space is not the full naive tensor product $\mathcal{H}_{\text{naive}}^{(n)}$ but the appropriate symmetry-projected

subspace:

$$\mathcal{H}_{\text{phys}}^{(n)} = \begin{cases} \text{Sym}^n \mathcal{H}_1 & (\text{bosons: symmetric tensor product}), \\ \Lambda^n \mathcal{H}_1 & (\text{fermions: antisymmetric tensor product}). \end{cases} \quad (15.63)$$

At this level of description — non-relativistic quantum mechanics with fixed particle number — the choice between (15.61) and (15.62) is an additional postulate: we impose it by hand based on the species of particle, without deriving it from any deeper principle.

The Pauli exclusion principle as a theorem. The antisymmetry condition (15.62) immediately implies the Pauli exclusion principle. If two fermions share the same quantum numbers — $x_i = x_j$ and $s_i = s_j$ — then exchanging their labels leaves the configuration unchanged, so the wavefunction must equal itself: $\Psi(\dots, x_i, s_i, \dots, x_i, s_i, \dots) = \Psi(\dots, x_i, s_i, \dots, x_i, s_i, \dots)$. But antisymmetry requires it to equal minus itself:

$$\Psi = -\Psi \implies \Psi = 0. \quad (15.64)$$

The wavefunction vanishes identically: no two fermions can occupy the same single-particle state. This is not an independent postulate; it is a consequence of antisymmetry, which is itself a consequence of the indistinguishability of identical particles. The stability of matter, the periodic table, neutron star structure, and the rigidity of ordinary solids all follow from this vanishing.

Why the choice of sign is not arbitrary: a forward pointer. At the non-relativistic level, the ± 1 choice appears to be a free parameter of the theory: we simply declare electrons to be fermions and photons to be bosons. This is unsatisfying. In fact the choice is not free: it is *determined by the spin of the particle*, and the determination is a theorem of relativistic quantum field theory — the *spin-statistics theorem*. The argument runs as follows:

- Relativistic causality requires that measurements at spacelike-separated points commute: $[\hat{O}(x), \hat{O}(y)] = 0$ for $(x - y)^2 < 0$.
- For integer-spin fields, this commutativity is maintained by the commuting (bosonic) quantization: $[a(\mathbf{k}), a^\dagger(\mathbf{k}')] = \delta^3(\mathbf{k} - \mathbf{k}')$.
- For half-integer-spin fields, commutativity would be violated by bosonic quantization but is maintained by the anticommuting (fermionic) quantization: $\{b(\mathbf{k}), b^\dagger(\mathbf{k}')\} = \delta^3(\mathbf{k} - \mathbf{k}')$.

The conclusion is forced: *integer spin $\Rightarrow$ bosons; half-integer spin $\Rightarrow$ fermions*. The ± 1 in (15.61)–(15.62) is not a postulate; it is a consequence of Lorentz invariance and causality. The spin-statistics theorem is one of the deepest results in theoretical physics: it connects the spacetime transformation properties of a field (its spin, encoded in the Lorentz representation) to the algebraic structure of its Fock space (whether its creation operators commute or anticommute). The non-relativistic treatment of this section must simply postulate the exchange symmetry; the relativistic treatment of the Dirac field reveals why it is forced.

How Fock space resolves the symmetrisation problem naturally. The symmetrisation requirement (15.63) appears as an additional constraint in the fixed- n formalism: one must project the naive tensor product onto the symmetric or antisymmetric subspace by hand. This is conceptually awkward and computationally cumbersome for large n . The Fock space formalism resolves this naturally: the creation operators $a^\dagger(\mathbf{k})$ (bosons) and $b^\dagger(\mathbf{k})$ (fermions) automatically produce (anti)symmetrised states. For bosons:

$$a^\dagger(\mathbf{k}_1)a^\dagger(\mathbf{k}_2)|0\rangle = a^\dagger(\mathbf{k}_2)a^\dagger(\mathbf{k}_1)|0\rangle, \quad (15.65)$$

because $[a^\dagger(\mathbf{k}_1), a^\dagger(\mathbf{k}_2)] = 0$: the two-particle state is automatically symmetric. For fermions:

$$b^\dagger(\mathbf{k}_1)b^\dagger(\mathbf{k}_2)|0\rangle = -b^\dagger(\mathbf{k}_2)b^\dagger(\mathbf{k}_1)|0\rangle, \quad (15.66)$$

because $\{b^\dagger(\mathbf{k}_1), b^\dagger(\mathbf{k}_2)\} = 0$: the two-particle state is automatically antisymmetric. And the Pauli exclusion principle:

$$b^\dagger(\mathbf{k})b^\dagger(\mathbf{k})|0\rangle = -b^\dagger(\mathbf{k})b^\dagger(\mathbf{k})|0\rangle \implies (b^\dagger(\mathbf{k}))^2|0\rangle = 0, \quad (15.67)$$

follows directly from $\{b^\dagger(\mathbf{k}), b^\dagger(\mathbf{k})\} = 2(b^\dagger(\mathbf{k}))^2 = 0$. The symmetrisation that had to be imposed by hand in the fixed- n formalism is *built into the algebra* of the Fock creation and annihilation operators. This is one of the deepest reasons why the Fock space formalism is more natural than the fixed- n formalism: the exchange symmetry of identical particles is not a constraint on the Hilbert space but an algebraic property of the operators that build it.

Summary of the three levels.

Framework	Symmetrisation	Status
Fixed- n (this section)	Imposed by hand: project onto $\text{Sym}^n \mathcal{H}$ or $\Lambda^n \mathcal{H}$	Additional postulate
Fock space	Built into algebra: $[a^\dagger, a^\dagger] = 0$ or $\{b^\dagger, b^\dagger\} = 0$	Consequence of operator algebra
Relativistic QFT	Forced by spin-statistics theorem: spin determines commutator vs anticommutator	Theorem from Lorentz invariance and causality

The fixed- n formalism exposes the problem of symmetrisation most nakedly: identical particles cannot be labelled, so the naive Hilbert space overcounts physical states by $n!$ (the number of permutations of n labels), and the overcounting must be corrected by projection. The Fock space formalism dissolves the problem by building the exchange symmetry into the creation operators. The relativistic theory explains why the exchange symmetry takes the specific form it does. Each level of the hierarchy is a more natural and complete treatment of the same underlying physical fact: identical particles are indistinguishable, and the quantum formalism must respect this indistinguishability at the level of its fundamental algebraic structure.

16 Non-Relativistic Quantum Field Theory: Varying Particle Number

In a GPT language model, the residual stream $h_t \in \mathbb{R}^d$ encodes extraordinarily rich semantics: syntactic role, factual associations, discourse context, and sentiment all coexist in a single vector. The vector is not a list of labelled slots; it is a superposition over directions in a high-dimensional space, each direction carrying a single interpretable concept. Sparse autoencoder research calls these directions *monosemantic*: each one encodes exactly one thing.

The Fock-space hidden state is built on exactly the same idea — and it is the only change needed to go from the n -particle quantum mechanics of the previous section to quantum field theory. The n -particle theory fixed the particle number at n ; the Fock space allows it to vary. The Schrödinger equation $\dot{h}_t = -iHh_t$ is unchanged. The Born rendering layer is unchanged. The only difference is the richness of the index that labels the basis vectors.

Throughout this section, bold letters $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{k} \in \mathbb{R}^3$ denote spatial three-vectors, not spacetime four-vectors, emphasising that the Hamiltonian formulation singles out time from the start.

16.1 The basis vectors of Fock space

A basis vector of Fock space is labelled by a complete specification of a particle configuration: a definite number n of particles, each with a definite momentum. We write:

$$|n; \mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n\rangle, \quad (16.1)$$

meaning: exactly n particles present, with momenta $\mathbf{k}_1, \dots, \mathbf{k}_n$. For $n = 0$ there is a unique basis vector, the *vacuum* $|0\rangle$: the state with no particles. These basis vectors are:

- **Mutually orthogonal:** $\langle n; \mathbf{k}_1, \dots | m; \mathbf{q}_1, \dots \rangle = 0$ whenever $n \neq m$ or any momenta differ.
- **Monosemantic:** each one encodes exactly one particle configuration — a definite number of particles with definite momenta, nothing else. This is the quantum analogue of a one-hot vector: the basis vector $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ is “hot” at the single configuration $(n, \mathbf{k}_1, \dots, \mathbf{k}_n)$ and zero everywhere else.
- **Complete:** every state in Fock space can be expressed as a superposition of these basis vectors.

For particles with spin or other internal quantum numbers, the basis vector carries additional labels: $|n; \mathbf{k}_1, s_1; \dots; \mathbf{k}_n, s_n\rangle$. For bosons, the basis vectors are symmetrised under exchange of any two labels; for fermions, antisymmetrised. As before, antisymmetry implies the Pauli exclusion principle: two fermionic particles cannot share the same quantum numbers.

16.2 The Fock hidden state as an explicit superposition

The general Fock hidden state is a superposition over all basis vectors (16.1), summing over all particle numbers and integrating over all momentum configurations:

$$\begin{aligned} h_t = & c_0(t) |0\rangle \\ & + \int \frac{d^3\mathbf{k}_1}{(2\pi)^3} c_1(\mathbf{k}_1; t) |1; \mathbf{k}_1\rangle \\ & + \int \frac{d^3\mathbf{k}_1}{(2\pi)^3} \frac{d^3\mathbf{k}_2}{(2\pi)^3} c_2(\mathbf{k}_1, \mathbf{k}_2; t) |2; \mathbf{k}_1, \mathbf{k}_2\rangle \\ & + \dots \end{aligned} \quad (16.2)$$

The coefficient $c_n(\mathbf{k}_1, \dots, \mathbf{k}_n; t)$ is the probability amplitude for finding exactly n particles with momenta $\mathbf{k}_1, \dots, \mathbf{k}_n$ at time t . The unit-norm condition $\|h_t\|^2 = 1$ becomes:

$$|c_0|^2 + \int \frac{d^3\mathbf{k}}{(2\pi)^3} |c_1(\mathbf{k})|^2 + \int \frac{d^3\mathbf{k}_1 d^3\mathbf{k}_2}{(2\pi)^6} |c_2(\mathbf{k}_1, \mathbf{k}_2)|^2 + \dots = 1. \quad (16.3)$$

The Born rule gives the probability of finding exactly n particles with momenta in a small cell around $(\mathbf{k}_1, \dots, \mathbf{k}_n)$ as $|c_n(\mathbf{k}_1, \dots, \mathbf{k}_n)|^2 \prod_j \frac{d^3\mathbf{k}_j}{(2\pi)^3}$.

The integral is a continuous sum. The integral $\int \frac{d^3\mathbf{k}}{(2\pi)^3}$ is simply a sum over all momentum values, where the momentum index is continuous rather than discrete. For a finite volume V , momenta are discrete and the integral becomes $\frac{1}{V} \sum_{\mathbf{k}}$. In the infinite-volume limit $V \rightarrow \infty$, the sum becomes the integral. There is no conceptual difference: both are linear superpositions over basis vectors.

16.3 Each basis vector is monosemantic: the Fock-SAE correspondence

In sparse autoencoder interpretations of GPT, a monosemantic direction f_k in the residual stream encodes exactly one concept: “France”, “past tense”, “toxic language”, and so on. The embedding $h = \sum_k \alpha_k f_k$ is a superposition of such directions, each weighted by its relevance to the current token.

The Fock basis vector $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ is monosemantic in exactly this sense: it encodes exactly one particle configuration, with no ambiguity. The coefficient $c_n(\mathbf{k}_1, \dots, \mathbf{k}_n; t)$ plays the role of α_k : it is the complex-valued amplitude with which h_t “attends to” that particular configuration. The superposition (16.2) is the quantum analogue of the sparse superposition of monosemantic features.

There are two differences from the GPT case. First, the weights c_n are complex rather than real: they carry phase information that determines how different basis vectors interfere under the Schrödinger evolution. Second, the superposition is exact and dense rather than sparse: the interaction Hamiltonian continuously redistributes amplitude among all particle configurations simultaneously.

The Fock hidden state is therefore polysemantic — it carries amplitude in infinitely many monosemantic basis vectors at once — exactly as the GPT residual stream is polysemantic, encoding many concepts simultaneously. The Born rendering layer is the unembedding: it projects h_t onto a chosen basis and extracts one classical outcome, exactly as the unembedding matrix of a transformer projects the residual stream onto the vocabulary.

16.4 Why this superposition belongs to Fock space: the direct sum

Fock space is the direct sum of all fixed-particle-number Hilbert spaces:

$$\mathcal{F} = \mathcal{H}^{(0)} \oplus \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)} \oplus \dots = \bigoplus_{n=0}^{\infty} \mathcal{H}^{(n)}. \quad (16.4)$$

Each $\mathcal{H}^{(n)}$ is spanned by all n -particle basis vectors $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$. Different sectors are orthogonal: $\mathcal{H}^{(n)} \perp \mathcal{H}^{(m)}$ for $n \neq m$. The n -particle component of the hidden state:

$$h_t^{(n)} = \int \prod_{j=1}^n \frac{d^3\mathbf{k}_j}{(2\pi)^3} c_n(\mathbf{k}_1, \dots, \mathbf{k}_n; t) |n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle \in \mathcal{H}^{(n)} \quad (16.5)$$

lives entirely within $\mathcal{H}^{(n)}$. The full Fock hidden state is their direct sum:

$$h_t = h_t^{(0)} \oplus h_t^{(1)} \oplus h_t^{(2)} \oplus \dots \in \mathcal{F}. \quad (16.6)$$

There is nothing more to Fock space than this. A Fock hidden state is a superposition of monosemantic basis vectors, each labelled by a particle configuration, with complex amplitudes c_n . The direct sum organises these basis vectors by particle number. There is no block structure to maintain, no special data structure required. The Fock hidden state is a vector in a Hilbert space, exactly as the GPT residual stream is a vector in $\mathbb{R}^d$. The only difference is the richness of the index.

16.5 Why momentum rather than position: the natural basis

In the n -particle theory we used both the position basis $\{|\mathbf{x}\rangle\}$ and the momentum basis $\{|\mathbf{k}\rangle\}$, related by the Fourier transform. For quantum field theory, the momentum basis is almost always preferred, for three reasons.

The free Hamiltonian $H_{\text{free}} = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \varepsilon_{\mathbf{k}} a^\dagger(\mathbf{k}) a(\mathbf{k})$ is diagonal in the momentum basis: each mode $\mathbf{k}$ has a definite energy $\varepsilon_{\mathbf{k}}$, making perturbation theory tractable. Interactions are local in position space but momentum-conserving in momentum space, which is the basic vertex structure of Feynman diagrams. And the Fock basis vectors $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ labelled by conserved momenta are the monosemantic eigenstates of the free dynamics: each one is a stationary state of H_{free} up to a phase.

16.6 The Schrödinger equation on Fock space

The Schrödinger equation:

$$\frac{d}{dt} h_t = -iH h_t, \quad h_t \in \mathcal{F}, \quad (16.7)$$

acts on $h_t \in \mathcal{F}$ in exactly the same way as on any Hilbert space. The free Hamiltonian H_{free} maps each basis vector $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ to a scalar multiple of itself (phase rotation at the total energy $\sum_j \varepsilon_{\mathbf{k}_j}$), preserving particle number. The interaction Hamiltonian H_{int} maps $|n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$ to a linear combination of basis vectors with different particle numbers: this is particle creation and annihilation. The coefficients $c_n(\mathbf{k}_1, \dots, \mathbf{k}_n; t)$ evolve continuously, with amplitude flowing between different particle configurations. The story of particle creation, propagation, scattering, and annihilation is the story of how the Schrödinger equation redistributes amplitude among monosemantic basis vectors over time.

QM (fixed particle number)	QFT (variable particle number)
Basis: $ \mathbf{k}\rangle$, one index	Basis: $ n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$, richer index
$h_t = \int \frac{d^3\mathbf{k}}{(2\pi)^3} c(\mathbf{k}; t) \mathbf{k}\rangle$	$h_t = \sum_n \int \prod_j \frac{d^3\mathbf{k}_j}{(2\pi)^3} c_n(\mathbf{k}_1, \dots; t) n; \mathbf{k}_1, \dots\rangle$
$\dot{h}_t = -iH h_t$	$\dot{h}_t = -iH h_t$ (unchanged)
H rotates amplitude among $ \mathbf{k}\rangle$	H also rotates amplitude among different n
Born: $P(\mathbf{k}) = c(\mathbf{k}) ^2$	Born: $P(n; \mathbf{k}_1, \dots) = c_n(\mathbf{k}_1, \dots) ^2$

The right column is obtained from the left by enriching the basis index. The equation $\dot{h}_t = -iH h_t$ is identical. Quantum field theory is not a new formalism; it is quantum mechanics with a richer set of basis vectors.

16.7 Creation and annihilation operators

The Fock sectors are connected by *creation* and *annihilation* operators, which are the fundamental building blocks of the Hamiltonian.

The creation operator $a^\dagger(\mathbf{k})$ adds one particle of momentum $\mathbf{k}$:

$$a^\dagger(\mathbf{k}) : \mathcal{H}^{(n)} \rightarrow \mathcal{H}^{(n+1)}, \quad a^\dagger(\mathbf{k}) |0\rangle = |\mathbf{k}\rangle. \quad (16.8)$$

The annihilation operator $a(\mathbf{k})$ removes one particle of momentum $\mathbf{k}$:

$$a(\mathbf{k}) : \mathcal{H}^{(n)} \rightarrow \mathcal{H}^{(n-1)}, \quad a(\mathbf{k}) |0\rangle = 0. \quad (16.9)$$

Canonical commutation and anticommutation relations. For bosons:

$$[a(\mathbf{k}), a^\dagger(\mathbf{k}')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'), \quad [a(\mathbf{k}), a(\mathbf{k}')] = 0. \quad (16.10)$$

For fermions:

$$\{a(\mathbf{k}), a^\dagger(\mathbf{k}')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'), \quad \{a(\mathbf{k}), a(\mathbf{k}')\} = 0. \quad (16.11)$$

These are consequences of the symmetry (bosons) or antisymmetry (fermions) of the Fock basis vectors. The fermionic relation $(a^\dagger(\mathbf{k}))^2 = 0$ is the Pauli exclusion principle in algebraic form.

The number operator $\hat{N} = \int \frac{d^3\mathbf{k}}{(2\pi)^3} a^\dagger(\mathbf{k})a(\mathbf{k})$ counts the total number of particles: $\hat{N} |n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle = n |n; \mathbf{k}_1, \dots, \mathbf{k}_n\rangle$.

16.8 The field operator and mode expansion

The *field operator* $\hat{\psi}(\mathbf{x})$ is the Fourier transform of the annihilation operators:

$$\hat{\psi}(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} a(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}, \quad \hat{\psi}^\dagger(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} a^\dagger(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}}. \quad (16.12)$$

The operator $\hat{\psi}(\mathbf{x})$ annihilates a particle at position $\mathbf{x}$; $\hat{\psi}^\dagger(\mathbf{x})$ creates one. The density operator $\hat{n}(\mathbf{x}) = \hat{\psi}^\dagger(\mathbf{x})\hat{\psi}(\mathbf{x})$ measures the local particle density.

Why promote Fourier coefficients to operators? In classical field theory, the Fourier decomposition $\psi(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \tilde{\psi}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}$ has ordinary complex numbers $\tilde{\psi}(\mathbf{k})$ as coefficients. Quantization promotes these to operators: $\tilde{\psi}(\mathbf{k}) \rightarrow a(\mathbf{k})$. Each Fourier mode $\mathbf{k}$ is an independent harmonic oscillator for a free field, and harmonic oscillator quantization replaces the classical amplitude with the ladder operators a and $a^\dagger$.

Crucially, we do not promote $\mathbf{x}$ or $\mathbf{k}$ to operators. They remain classical labels — coordinates and frequencies. What we promote is the field amplitude at each mode: the Fourier coefficient $\tilde{\psi}(\mathbf{k})$ becomes the operator $a(\mathbf{k})$ on the Fock space. The position $\mathbf{x}$ is a spatial coordinate, not a dynamical degree of freedom.

Time and space separation. The field operator $\hat{\psi}(\mathbf{x})$ at fixed time t is a function of the spatial coordinate $\mathbf{x}$ only. Time has been singled out: the Fock basis vectors $|n; \mathbf{k}_1, \dots\rangle$ are spatial snapshots, and the Schrödinger equation propagates h_t forward in time. The time-dependent Heisenberg-picture field:

$$\hat{\psi}(\mathbf{x}, t) = e^{iHt} \hat{\psi}(\mathbf{x}) e^{-iHt} = \int \frac{d^3\mathbf{k}}{(2\pi)^3} a(\mathbf{k}) e^{i(\mathbf{k}\cdot\mathbf{x} - \varepsilon_{\mathbf{k}}t)} \quad (16.13)$$

rotates each momentum mode at its natural frequency $\varepsilon_{\mathbf{k}}$. Lorentz invariance is not manifest in this formulation — time and space are treated differently — but it can be restored by requiring the Lagrangian to be a Lorentz scalar.

16.9 The Hamiltonian motor as a polynomial of field operators

The Hamiltonian is written directly in terms of creation and annihilation operators. For non-relativistic particles of mass m :

Free Hamiltonian:

$$H_{\text{free}} = \int d^3\mathbf{x} \, \hat{\psi}^\dagger(\mathbf{x}) \left(-\frac{\nabla^2}{2m} \right) \hat{\psi}(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \varepsilon_{\mathbf{k}} a^\dagger(\mathbf{k}) a(\mathbf{k}), \quad (16.14)$$

where $\varepsilon_{\mathbf{k}} = \mathbf{k}^2/2m$ is the free particle dispersion relation. This is the non-relativistic Hamiltonian motor: a polynomial of degree 2 in the creation and annihilation operators, diagonal in the momentum-space Fock basis.

Two-body interaction:

$$H_{\text{int}} = \frac{1}{2} \int d^3\mathbf{x} d^3\mathbf{y} \, \hat{\psi}^\dagger(\mathbf{x}) \hat{\psi}^\dagger(\mathbf{y}) V(\mathbf{x} - \mathbf{y}) \hat{\psi}(\mathbf{y}) \hat{\psi}(\mathbf{x}). \quad (16.15)$$

In momentum space, substituting the mode expansion $\hat{\psi}(\mathbf{x}) = \int \frac{d^3k}{(2\pi)^3} a(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}$:

$$H_{\text{int}} = \frac{1}{2} \int \frac{d^3k_1 d^3k_2 d^3q}{(2\pi)^9} \tilde{V}(\mathbf{q}) a^\dagger(\mathbf{k}_1 + \mathbf{q}) a^\dagger(\mathbf{k}_2 - \mathbf{q}) a(\mathbf{k}_2) a(\mathbf{k}_1), \quad (16.16)$$

where $\tilde{V}(\mathbf{q}) = \int d^3r V(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}}$ is the Fourier transform of the interaction potential. This is a polynomial of degree 4: two creation operators and two annihilation operators at each vertex. The vertex describes a specific process: two particles with momenta $\mathbf{k}_1$ and $\mathbf{k}_2$ exchange momentum $\mathbf{q}$ via the potential $\tilde{V}(\mathbf{q})$, emerging with momenta $\mathbf{k}_1 + \mathbf{q}$ and $\mathbf{k}_2 - \mathbf{q}$. Total momentum is conserved at each vertex: $(\mathbf{k}_1 + \mathbf{q}) + (\mathbf{k}_2 - \mathbf{q}) = \mathbf{k}_1 + \mathbf{k}_2$.

The full Hamiltonian is:

$$H = \underbrace{\int \frac{d^3k}{(2\pi)^3} \varepsilon_{\mathbf{k}} a^\dagger(\mathbf{k}) a(\mathbf{k})}_{\text{degree 2: free motor}} + \underbrace{\frac{1}{2} \int \frac{d^3k_1 d^3k_2 d^3q}{(2\pi)^9} \tilde{V}(\mathbf{q}) a^\dagger(\mathbf{k}_1 + \mathbf{q}) a^\dagger(\mathbf{k}_2 - \mathbf{q}) a(\mathbf{k}_2) a(\mathbf{k}_1)}_{\text{degree 4: interaction motor}}. \quad (16.17)$$

This is the non-relativistic Hamiltonian motor: a polynomial of degree 4 in creation and annihilation operators, acting on the Fock space $\mathcal{F}$. The degree-2 part is diagonal (phase rotation). The degree-4 part is off-diagonal (scattering).

In condensed matter physics, this Hamiltonian is often written down directly, without ever passing through a classical Lagrangian. This is the correct conceptual order: specify the Fock space, write the Hamiltonian polynomial, let the Schrödinger equation evolve h_t . The Lagrangian and path integral are derived objects.

16.9.1 The eigenstructure of the free theory

The eigenstates of H_{free} are the Fock states: states of definite particle number and definite momentum for each particle.

One-particle sector. A single particle with momentum $\mathbf{k}$:

$$|\mathbf{k}\rangle = a^\dagger(\mathbf{k}) |0\rangle, \quad H_{\text{free}} |\mathbf{k}\rangle = \varepsilon_{\mathbf{k}} |\mathbf{k}\rangle = \frac{\mathbf{k}^2}{2m} |\mathbf{k}\rangle. \quad (16.18)$$

The dispersion relation $\varepsilon_{\mathbf{k}} = k^2/2m$ is parabolic: the non-relativistic kinetic energy. The eigenstates are labelled by continuous momentum $\mathbf{k} \in \mathbb{R}^3$; the spectrum is continuous and non-degenerate (given $\mathbf{k}$). For fermions with spin, the label includes the spin projection: $|\mathbf{k}, \sigma\rangle$, $\sigma = \uparrow, \downarrow$.

Two-particle sector. Two particles with momenta $\mathbf{k}_1$ and $\mathbf{k}_2$:

$$|\mathbf{k}_1, \mathbf{k}_2\rangle = a^\dagger(\mathbf{k}_1) a^\dagger(\mathbf{k}_2) |0\rangle, \quad H_{\text{free}} |\mathbf{k}_1, \mathbf{k}_2\rangle = (\varepsilon_{\mathbf{k}_1} + \varepsilon_{\mathbf{k}_2}) |\mathbf{k}_1, \mathbf{k}_2\rangle. \quad (16.19)$$

The energy is the sum of individual kinetic energies. For bosons, $|\mathbf{k}_1, \mathbf{k}_2\rangle = |\mathbf{k}_2, \mathbf{k}_1\rangle$ (symmetric). For fermions, $|\mathbf{k}_1, \mathbf{k}_2\rangle = -|\mathbf{k}_2, \mathbf{k}_1\rangle$ (antisymmetric), and $\mathbf{k}_1 = \mathbf{k}_2$ (with same spin) gives zero (Pauli exclusion).

The degeneracy at fixed total energy E is the set of all momentum pairs $(\mathbf{k}_1, \mathbf{k}_2)$ with $k_1^2/2m + k_2^2/2m = E$: a five-dimensional surface in the six-dimensional space $(\mathbf{k}_1, \mathbf{k}_2)$. The degeneracy is continuous and enormous.

n -particle sector. The general free Fock state:

$$|\mathbf{k}_1, \dots, \mathbf{k}_n\rangle = a^\dagger(\mathbf{k}_1) \cdots a^\dagger(\mathbf{k}_n) |0\rangle, \quad E = \sum_{j=1}^n \varepsilon_{\mathbf{k}_j} = \sum_{j=1}^n \frac{k_j^2}{2m}. \quad (16.20)$$

The motor is diagonal: it assigns each Fock state the sum of its constituent kinetic energies. No scattering, no correlations, no entanglement. Each particle moves independently, at constant velocity, forever.

The Fermi gas ground state. For N non-interacting fermions (spin- $\frac{1}{2}$) in a box of volume V , the ground state fills the lowest-energy single-particle states up to the **Fermi momentum** k_F :

$$|\text{FS}\rangle = \prod_{\substack{|\mathbf{k}| < k_F \\ \sigma=\uparrow, \downarrow}} a_\sigma^\dagger(\mathbf{k}) |0\rangle, \quad (16.21)$$

where the Fermi momentum is determined by the particle density $n = N/V$:

$$k_F = (3\pi^2 n)^{1/3}. \quad (16.22)$$

Every single-particle state with $|\mathbf{k}| < k_F$ is occupied (both spins); every state with $|\mathbf{k}| > k_F$ is empty. The occupied states fill a sphere of radius k_F in momentum space: the **Fermi sphere**. The surface of the sphere ($|\mathbf{k}| = k_F$) is the **Fermi surface**.

The Fermi energy is:

$$E_F = \varepsilon_{k_F} = \frac{k_F^2}{2m} = \frac{(3\pi^2 n)^{2/3}}{2m}. \quad (16.23)$$

For electrons in a metal: $n \sim 10^{28} \text{ m}^{-3}$, $E_F \sim 5 \text{ eV}$, $T_F = E_F/k_B \sim 60,000 \text{ K}$. The Fermi energy is enormous compared to room temperature ($k_B T \sim 0.025 \text{ eV}$): the electron gas is deeply degenerate. Quantum mechanics (the Pauli exclusion principle) dominates the structure of the ground state.

The total ground state energy is:

$$E_{\text{gs}} = 2 \sum_{|\mathbf{k}| < k_F} \frac{k^2}{2m} = V \cdot \frac{3}{5} n E_F. \quad (16.24)$$

The factor $3/5$ is the average over the Fermi sphere: $\langle k^2 \rangle = (3/5)k_F^2$.

This is the non-interacting ground state: a product state (a single Slater determinant), with no correlations, no entanglement between particles, no collective behaviour. It is the free-field analogue of the harmonic oscillator ground state $|0\rangle$ —the state annihilated by all lowering operators. In this case, the “lowering” of the Fermi sea is blocked by the Pauli principle: you cannot remove a particle from below the Fermi surface and place it below itself.

16.9.2 How interactions change the eigenstructure

The interaction H_{int} (degree 4) couples different Fock states: it scatters pairs of particles between momentum states, mixing different free Fock vectors. The eigenstates of the full $H = H_{\text{free}} + H_{\text{int}}$ are not free Fock states. They are superpositions of free Fock states—dressed, correlated, entangled.

The interacting ground state. The ground state of the interacting system $|\Omega\rangle$ is not the Fermi sea $|\text{FS}\rangle$. It is a superposition:

$$|\Omega\rangle = C_0 |\text{FS}\rangle + \sum_{\substack{\mathbf{k}_1, \mathbf{k}_2 < k_F \\ \mathbf{k}_3, \mathbf{k}_4 > k_F}} C_{\mathbf{k}_1 \mathbf{k}_2}^{\mathbf{k}_3 \mathbf{k}_4} a^\dagger(\mathbf{k}_3) a^\dagger(\mathbf{k}_4) a(\mathbf{k}_2) a(\mathbf{k}_1) |\text{FS}\rangle + \dots \quad (16.25)$$

The leading correction involves **particle-hole excitations**: two particles below the Fermi surface ($\mathbf{k}_1, \mathbf{k}_2 < k_F$) are scattered to states above the Fermi surface ($\mathbf{k}_3, \mathbf{k}_4 > k_F$) by the interaction, leaving two holes below. The coefficients $C_{\mathbf{k}_1 \mathbf{k}_2}^{\mathbf{k}_3 \mathbf{k}_4}$ are proportional to $\tilde{V}(\mathbf{q})/(E_{\text{excited}} - E_{\text{FS}})$: the interaction matrix element divided by the energy denominator.

The interacting ground state has *correlations*: the positions of the particles are entangled by the interaction. The particles are no longer independent. They avoid each other (for repulsive interactions) or cluster together (for attractive interactions), in patterns encoded in the coefficients C .

Quasiparticles: Landau's Fermi liquid theory. Landau's central insight (1956–1958): the low-energy excitations of the interacting Fermi system are not free particles but **quasiparticles**—dressed excitations that behave like particles with renormalised properties.

A quasiparticle is created by adding a particle just above the Fermi surface:

$$|\mathbf{k}, \text{qp}\rangle = a^\dagger(\mathbf{k}) |\Omega\rangle + (\text{corrections from } H_{\text{int}}). \quad (16.26)$$

The bare particle $a^\dagger(\mathbf{k}) |\Omega\rangle$ is not an eigenstate of H : the interaction immediately scatters it, creating particle-hole pairs. But for $|\mathbf{k}|$ near k_F , the dressed state—the quasiparticle—is *approximately* an eigenstate, with a well-defined energy and a long (but finite) lifetime.

The quasiparticle has:

- **Renormalised dispersion:** $\tilde{\varepsilon}_{\mathbf{k}} = \varepsilon_{\mathbf{k}} + \Sigma(\mathbf{k}, \tilde{\varepsilon}_{\mathbf{k}})$, where Σ is the self-energy—the energy shift from the interaction with the surrounding particles. Near the Fermi surface, the dispersion is approximately linear: $\tilde{\varepsilon}_{\mathbf{k}} \approx v_F^*(|\mathbf{k}| - k_F)$, where $v_F^* = k_F/m^*$ defines the effective mass m^* .
- **Effective mass:** $m^* \neq m$. The quasiparticle moves through the interacting medium as if it had a different mass. For electrons in metals: m^*/m ranges from ~ 1 (simple metals like sodium) to ~ 1000 (heavy-fermion compounds like CeAl_3).
- **Finite lifetime:** $\tau_{\mathbf{k}} \propto 1/(\varepsilon_{\mathbf{k}} - E_F)^2$. The quasiparticle decays by scattering into particle-hole pairs. The lifetime diverges as $\varepsilon_{\mathbf{k}} \rightarrow E_F$: quasiparticles at the Fermi surface are infinitely long-lived. This is why the quasiparticle concept works: near the Fermi surface, the dressed excitations are sharp (long-lived), and the interacting system looks like a gas of weakly interacting quasiparticles.
- **Quasiparticle residue:** $Z_{\mathbf{k}} = |\langle \mathbf{k}, \text{qp} | a^\dagger(\mathbf{k}) | \Omega \rangle|^2 < 1$. The overlap between the bare particle and the quasiparticle is less than 1: the quasiparticle is partly a bare particle and partly a cloud of particle-hole excitations. $Z = 1$ means no dressing (free particle); $Z = 0$ means the quasiparticle has dissolved entirely into the many-body continuum.

The Fermi liquid theory states that, for a wide class of interacting fermion systems, the low-energy eigenstructure of the interacting Hamiltonian is in one-to-one correspondence with the free eigenstructure:

Free theory	Interacting theory (Fermi liquid)
Particle $a^\dagger(\mathbf{k}) 0\rangle$	Quasiparticle $a^\dagger(\mathbf{k}) \Omega\rangle + \dots$
Dispersion $\varepsilon_{\mathbf{k}} = k^2/2m$	Renormalised dispersion $\tilde{\varepsilon}_{\mathbf{k}}$, effective mass m^*
Fermi surface at k_F	Fermi surface at k_F (Luttinger's theorem: same k_F)
Infinite lifetime	$\tau \propto 1/(E - E_F)^2$ (long near E_F)
No correlations	Correlations encoded in Landau parameters $f(\mathbf{k}, \mathbf{k}')$
$Z = 1$	$0 < Z < 1$ (spectral weight reduced)

This one-to-one correspondence is called **adiabatic continuity**: if you turn on the interaction slowly (from $V = 0$ to the physical V), each free eigenstate evolves continuously into an interacting eigenstate, without any level crossing or phase transition. The labels (momentum, spin) survive; the properties (energy, lifetime, weight) are renormalised.

When Fermi liquid theory fails. Adiabatic continuity fails when the interaction drives a phase transition: a qualitative change in the ground state that cannot be reached by smoothly deforming the free Fermi sea. Examples:

- **Superconductivity:** an attractive interaction (mediated by phonons) causes electrons near the Fermi surface to form Cooper pairs. The ground state is a condensate of pairs—a BCS state—qualitatively different from the Fermi sea. The quasiparticle picture breaks down: the low-energy excitations are Bogoliubov quasiparticles (superpositions of particles and holes) with a gap Δ in the spectrum.
- **Magnetism:** a repulsive interaction can drive a Stoner instability, splitting the Fermi surface into spin-up and spin-down surfaces with different radii. The ground state has a net magnetisation—a broken symmetry not present in the Fermi sea.
- **Mott insulator:** for electrons on a lattice with strong on-site repulsion U , the Fermi liquid description fails entirely. The ground state is an insulator (no conduction), even though band theory (the non-interacting eigenstructure) predicts a metal. The interaction localises the electrons on lattice sites, destroying the quasiparticle concept.
- **Non-Fermi liquids:** in some systems (e.g. electrons in one dimension, certain quantum critical points), the quasiparticle residue $Z \rightarrow 0$: the bare particle has zero overlap with the interacting eigenstates. The interacting eigenstructure bears no resemblance to the free eigenstructure, and no perturbative dressing can bridge the gap.

These failures are the most interesting phenomena in condensed matter physics: they are the cases where the interaction does not merely renormalise the free eigenstructure but *reorganises* it qualitatively. The Hubbard-Stratonovich transformation (Section ??) provides a natural framework for these reorganisations: the auxiliary field (the order parameter of the new phase) captures the collective degree of freedom that the interaction generates.

16.9.3 Specific examples of the non-relativistic motor

The electron gas (jellium). N electrons in a uniform positive background (the “jelly” of positive charge that neutralises the system). The interaction is the Coulomb potential

$V(\mathbf{r}) = e^2/|\mathbf{r}|$, with Fourier transform $\tilde{V}(\mathbf{q}) = 4\pi e^2/q^2$. The motor is:

$$H_{\text{jellium}} = \sum_{\mathbf{k},\sigma} \frac{k^2}{2m} a_{\mathbf{k}\sigma}^\dagger a_{\mathbf{k}\sigma} + \frac{1}{2V} \sum_{\mathbf{q} \neq 0} \frac{4\pi e^2}{q^2} \rho_{\mathbf{q}} \rho_{-\mathbf{q}}, \quad (16.27)$$

where $\rho_{\mathbf{q}} = \sum_{\mathbf{k},\sigma} a_{\mathbf{k}+\mathbf{q},\sigma}^\dagger a_{\mathbf{k},\sigma}$ is the density operator. The $\mathbf{q} = 0$ term is excluded (it cancels against the positive background). This is the Hamiltonian that Hubbard applied the HST to in 1959 (Section ??): the $1/q^2$ Coulomb vertex is the interaction that drives screening, plasmons, and the physics of metals.

The dimensionless coupling constant is $r_s = r_0/a_0$, where $r_0 = (3/4\pi n)^{1/3}$ is the mean inter-particle spacing and $a_0 = \hbar^2/me^2$ is the Bohr radius. For $r_s \ll 1$ (high density): the kinetic energy dominates, the system is a weakly interacting Fermi liquid, and perturbation theory (the random-phase approximation) works. For $r_s \gg 1$ (low density): the interaction dominates, the system crystallises into a Wigner crystal, and the Fermi liquid description fails.

Contact interaction (cold atoms). Ultracold atomic gases interact via a short-range potential that, at low energies, is fully characterised by a single parameter: the s -wave scattering length a_s . The effective interaction is a contact potential:

$$V(\mathbf{r}) = g \delta^3(\mathbf{r}), \quad g = \frac{4\pi \hbar^2 a_s}{m}, \quad \tilde{V}(\mathbf{q}) = g. \quad (16.28)$$

The Fourier transform is a constant: the interaction vertex has no momentum dependence. The motor is:

$$H_{\text{cold}} = \sum_{\mathbf{k}} \frac{k^2}{2m} a_{\mathbf{k}}^\dagger a_{\mathbf{k}} + \frac{g}{2V} \sum_{\mathbf{k}_1, \mathbf{k}_2, \mathbf{q}} a_{\mathbf{k}_1+\mathbf{q}}^\dagger a_{\mathbf{k}_2-\mathbf{q}}^\dagger a_{\mathbf{k}_2} a_{\mathbf{k}_1}. \quad (16.29)$$

For bosons: this motor, at sufficiently low temperature, produces the Bose-Einstein condensate—macroscopic occupation of the $\mathbf{k} = 0$ mode, with the Bogoliubov excitation spectrum $E_k = \sqrt{\varepsilon_k(\varepsilon_k + 2gn)}$ replacing the free dispersion $\varepsilon_k = k^2/2m$. For fermions with attractive interaction ($g < 0$): Cooper pairing and BCS superfluidity.

The Hubbard model (lattice). On a lattice with L sites, the creation operators are labelled by site index i and spin σ : $c_{i\sigma}^\dagger$. The Hubbard model is:

$$H_{\text{Hubbard}} = -t \sum_{\langle ij \rangle, \sigma} c_{i\sigma}^\dagger c_{j\sigma} + U \sum_i \hat{n}_{i\uparrow} \hat{n}_{i\downarrow}, \quad (16.30)$$

where t is the hopping amplitude (the one-body term: a particle hops from site j to neighbouring site i), and U is the on-site repulsion (the two-body term: two electrons on the same site pay an energy cost U). The motor is degree 2 (hopping) + degree 4 (interaction) = degree 4.

The Hubbard model is the minimal motor that captures the competition between kinetic energy (which delocalises particles, favouring a metallic Fermi liquid) and interaction energy (which localises particles, favouring a Mott insulator or magnetic order). Its eigenstructure—as a function of the ratio U/t , the filling $n = N/L$, and the lattice geometry—contains Fermi liquids, antiferromagnets, Mott insulators, and (it is believed but not proved) d -wave superconductors. The full solution of the Hubbard model in two dimensions is one of the great open problems of condensed matter physics.

16.9.4 The motor that condensed matter physics writes directly

In condensed matter physics, the Hamiltonian (16.17) is often written down directly—not derived from a classical Lagrangian, not obtained by quantising a classical field, but *postulated*

as the starting point of the theory. The physicist specifies: the Fock space (which creation and annihilation operators exist, with which commutation or anticommutation relations), the free dispersion (the single-particle energies $\varepsilon_{\mathbf{k}}$), and the interaction polynomial (the vertex $\tilde{V}(\mathbf{q})$ or the Hubbard U). The Schrödinger equation $\dot{h}_t = -iHh_t$ then evolves the Fock hidden state. The Born rule projects it onto observables.

This is the correct conceptual order: the game engine is a polynomial motor on Fock space, specified algebraically. The Lagrangian, the path integral, the classical field equation—these are derived objects, useful for calculations but not fundamental. The fundamental object is the Hamiltonian polynomial (16.17): a finite list of monomials in a and $a^\dagger$, each with a coefficient, acting on the Fock space at every tick.

The eigenstructure of this motor—free Fock states for H_{free} , quasiparticles and collective modes for $H_{\text{free}} + H_{\text{int}}$, phase transitions when adiabatic continuity fails—is the entire content of condensed matter physics. The exponential dimension of the Fock space ($\binom{M}{N}$ for N fermions in M orbitals, or $\binom{M+N-1}{N}$ for bosons) is the exponential abyss: the gap between the hidden state h_t and the Born-rendered eigenvalue that the observer receives. The observer sees a resistivity, a magnetic susceptibility, a specific heat—a single number projected from a space of dimension $10^{10^{23}}$.

16.10 Second quantization: what the name really means

The framework just described is called *second quantization* in the literature. The name is misleading and should be understood correctly.

It is not a second application of quantization rules. It is not the quantization of the field $\hat{\psi}(\mathbf{x})$ as a classical object. There is no “first quantization” of a classical field followed by a “second quantization” of the wavefunction.

Second quantization is the extension of the Hilbert space from the fixed- n spaces $\mathcal{H}^{(n)}$ to the full Fock space $\mathcal{F} = \bigoplus_n \mathcal{H}^{(n)}$. The “second” refers to the second level of description: the first level is the single-particle Hilbert space $\mathcal{H}^{(1)}$; the second level is the Fock space built from it by direct sum. A better name would be *Fock space quantization* or *variable-particle-number quantization*. We use “second quantization” only to connect to the standard literature, while insisting that it is simply the direct-sum construction, no more mysterious than (16.4).

16.11 Physical examples: the electron gas and phonons

The electron gas. Consider electrons (fermions, spin- $\frac{1}{2}$) in a metal, interacting via the Coulomb potential. The Fock basis vectors carry momentum $\mathbf{k}$ and spin $s \in \{\uparrow, \downarrow\}$. The Hamiltonian:

$$H = \sum_s \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{\mathbf{k}^2}{2m} b_s^\dagger(\mathbf{k}) b_s(\mathbf{k}) + \frac{e^2}{2} \int \frac{d^3\mathbf{q}}{(2\pi)^3} \frac{4\pi}{\mathbf{q}^2} \sum_{s,s'} \int \frac{d^3\mathbf{k}_1 d^3\mathbf{k}_2}{(2\pi)^6} b_s^\dagger(\mathbf{k}_1 + \mathbf{q}) b_{s'}^\dagger(\mathbf{k}_2 - \mathbf{q}) b_{s'}(\mathbf{k}_2) b_s(\mathbf{k}_1) \quad (16.31)$$

is a polynomial of fermionic creation and annihilation operators. The ground state $|\Omega\rangle$ is the Fermi sea: all states below the Fermi momentum k_F are occupied. Quasiparticle excitations near k_F — electron-like and hole-like — are the emergent degrees of freedom of Fermi liquid theory.

Phonons: emergent bosons from a quantum lattice. The atoms in a crystal oscillate around their equilibrium positions. In the harmonic approximation, Fourier-transforming to normal modes diagonalises the Hamiltonian:

$$H_{\text{phonon}} = \sum_{\mathbf{k}} \omega(\mathbf{k}) \left(b^\dagger(\mathbf{k}) b(\mathbf{k}) + \frac{1}{2} \right), \quad (16.32)$$

where $b^\dagger(\mathbf{k})$ and $b(\mathbf{k})$ satisfy bosonic commutation relations. The phonon is an emergent boson: a collective excitation of the lattice that inherits its quantum nature from the underlying atomic commutation relations $[\hat{u}_i, \hat{p}_j] = i\delta_{ij}$. When the atom-displacement operators are quantized, the normal-mode operators inherit $[b, b^\dagger] = 1$. The phonon has discrete energy levels $\omega(\mathbf{k})(n + \frac{1}{2})$, zero-point energy, and Bose-Einstein statistics — all consequences of the quantum nature of the atoms it is built from.

In the Fock framework, the phonon is treated exactly like a fundamental particle: it has creation and annihilation operators, a Fock space $\mathcal{F}_{\text{phonon}}$, and a Hamiltonian (16.32). The Born rendering of the phonon field is the dynamic structure factor:

$$S(\mathbf{q}, \omega) = \int dt e^{i\omega t} \langle \hat{n}(\mathbf{q}, t) \hat{n}(-\mathbf{q}, 0) \rangle_\beta, \quad (16.33)$$

a thermal average (Level 3 Born rendering) of a collective density observable. The phonon appears as a sharp peak at $\omega = \omega(\mathbf{q})$: its quantum nature is inferred from the peak shape and temperature dependence, not from directly observing a phonon Fock state.

16.12 Condensate and spontaneous symmetry breaking

A *condensate* forms when a macroscopic number of bosons occupy the same single-particle state. In the degeneracy framework, it is the cleanest example of a Born observable crossing from quantum random (Level 1) to deterministic classical (Level 4).

Why the zero-momentum mode

The single-particle states of a free Bose gas are labelled by momentum $\mathbf{k}$, with energy $\varepsilon_{\mathbf{k}} = \mathbf{k}^2/2m$. The Bose-Einstein distribution:

$$\langle \hat{n}_{\mathbf{k}} \rangle = \frac{1}{e^{\beta(\varepsilon_{\mathbf{k}} - \mu)} - 1} \quad (16.34)$$

gives $O(1)$ occupation for $\mathbf{k} \neq \mathbf{0}$ (the energy keeps the denominator bounded), but for $\mathbf{k} = \mathbf{0}$, $\varepsilon_{\mathbf{0}} = 0$ and the occupation:

$$\langle \hat{n}_{\mathbf{0}} \rangle = \frac{1}{e^{-\beta\mu} - 1} \xrightarrow{\mu \rightarrow 0^-} \infty \quad (16.35)$$

diverges. In a finite system with N particles, the zero-momentum mode saturates at $N_0 = O(N)$. The mode is singled out by having the lowest energy: it is the cheapest mode for bosons to pile into, and Bose-Einstein statistics (which enhance rather than forbid multiple occupancy) allow unlimited piling.

The condensate:

$$\langle \hat{n}_{\mathbf{0}} \rangle = N_0 \sim O(N) \gg 1, \quad \langle \hat{n}_{\mathbf{k}} \rangle \sim O(1) \text{ for } \mathbf{k} \neq \mathbf{0}. \quad (16.36)$$

The Born observable and its degeneracy

The key observable is the *number operator for the zero-momentum mode*:

$$\hat{n}_{\mathbf{0}} = a^\dagger(\mathbf{0}) a(\mathbf{0}). \quad (16.37)$$

Its eigenvalues are $n_0 = 0, 1, 2, \dots$. In the full Fock space $\mathcal{F} = \mathcal{F}_{\mathbf{0}} \otimes \mathcal{F}_{\text{other}}$, each eigenvalue n_0 has *enormous degeneracy*: every configuration of all other modes ($\mathbf{k} \neq \mathbf{0}$) shares the same eigenvalue. The observable submitted to the Born layer is:

$$\hat{n}_{\mathbf{0}} \otimes \mathbf{1}_{\text{other modes}}. \quad (16.38)$$

The identity on all other modes provides the degeneracy: $d_{n_0} = \dim \mathcal{F}_{\text{other}}$ —astronomically large.

This is the tensor product specialisation from the entanglement section: the “system” is the $\mathbf{k} = \mathbf{0}$ mode; the “environment” is all other modes. The identity on the environment is the source of the degeneracy.

The projected superposition

The ground state $|\Omega\rangle$ is predominantly a superposition over Fock basis vectors with $n_0 \approx N_0$:

$$|\Omega\rangle \approx \sum_{n_0 \approx N_0} c_{n_0} |n_0; \mathbf{k} = \mathbf{0}\rangle \otimes |\text{other modes}\rangle. \quad (16.39)$$

The projected superposition onto eigenvalue n_0 :

$$h_t^{(n_0)} = c_{n_0} |n_0; \mathbf{k} = \mathbf{0}\rangle \otimes \sum_{\text{other}} c_{\text{other}}^{(n_0)} |\text{other}\rangle, \quad (16.40)$$

retains the full superposition over all other modes (the OR sum over the degenerate index).

Variance and entropy: the transition to classical

The Born probability $P(n_0) = \|h_t^{(n_0)}\|^2$ is sharply peaked around N_0 . The variance:

$$\frac{\text{Var}(\hat{n}_0)}{N_0^2} \sim \frac{N_0}{N_0^2} = \frac{1}{N_0} \sim \frac{1}{N} \rightarrow 0. \quad (16.41)$$

The entropy of the Born distribution over n_0 :

$$S_{\text{Born}}(\hat{n}_0) = - \sum_{n_0} P(n_0) \log P(n_0) \sim \frac{1}{2} \log N_0 \quad (16.42)$$

grows only logarithmically with N_0 , while $\log \dim \mathcal{H}$ grows as N or faster. The ratio $S_{\text{Born}}/\log \dim \mathcal{H} \rightarrow 0$: the Born distribution becomes a delta function relative to the full Hilbert space.

Compare the condensed and non-condensed modes:

Mode	$\langle \hat{n} \rangle$	Rel. variance	Entropy	Born level
$\mathbf{k} \neq \mathbf{0}$ (non-condensed)	$O(1)$	$O(1)$	$O(1)$	Level 1 (quantum)
$\mathbf{k} = \mathbf{0}$, $T < T_c$ (condensed)	$N_0 \sim N$	$1/N \rightarrow 0$	$\sim \frac{1}{2} \log N$	Level 4 (classical)
Any mode, $T > T_c$	$O(1)$	$O(1)$	$O(1)$	Level 1 (quantum)

The condensate is the unique mode for which the Born rendering crosses from quantum random to deterministic classical. The crossing is controlled by the occupation N_0 : when $N_0 \gg 1$, the variance vanishes and the Born output is effectively certain.

The classical field from the Born rendering

The operator $a(\mathbf{0})$ has relative quantum fluctuation $\Delta a/\langle a \rangle \sim 1/\sqrt{N_0} \rightarrow 0$. In the thermodynamic limit, $a(\mathbf{0})$ can be replaced by the classical number $\sqrt{N_0} e^{i\theta}$, and the order parameter:

$$\psi_0 = \langle \Omega | \hat{\psi}(\mathbf{x}) | \Omega \rangle = \sqrt{\frac{N_0}{V}} e^{i\theta} \quad (16.43)$$

is a deterministic classical complex number—a Born Level 4 rendering of the field operator.

Spontaneous symmetry breaking as a Born rendering statement

The order parameter (16.43) has a definite phase θ : the $U(1)$ symmetry $\psi \rightarrow e^{i\alpha}\psi$ is spontaneously broken. In the degeneracy framework, this is a statement about the Born observable, not about the hidden state.

The Hamiltonian conserves total particle number ($[H, \hat{N}] = 0$), so the hidden state h_t is a superposition of fixed- N sectors, each $U(1)$ -invariant. But the field operator $\hat{\psi}(\mathbf{x})$ —the observable whose Born rendering produces the order parameter —selects a definite phase when the Born distribution becomes deterministic ($N_0 \rightarrow \infty$).

The broken symmetry is in the rendering, not in the hidden state. The classical order parameter has a definite phase; the quantum hidden state does not. This is the general principle: spontaneous symmetry breaking is the Born rendering of a symmetric hidden state through an observable whose Level 4 output picks a definite value from a symmetric manifold.

Bose–Einstein condensation and superfluidity

Below T_c , the quasiparticle excitations above the condensate —Bogoliubov phonons—acquire a linear dispersion $\omega \propto |\mathbf{k}|$ at small $\mathbf{k}$, replacing the quadratic $\omega \propto \mathbf{k}^2$ of free particles. By Landau’s criterion, a moving object cannot create excitations below the minimum phase velocity $v_c = \min(\omega/|\mathbf{k}|)$ and experiences no drag: superfluidity.

The observable here is the drag force on the moving object—a macroscopic, highly degenerate observable. Its Born rendering is deterministic (Level 4) and gives zero below v_c : a classical, dissipation-free flow.

Cooper pairs and superconductivity

In a superconductor, pairs of electrons with opposite momenta and spins form composite bosons—Cooper pairs—of charge $2e$ and spin 0. The pair condensate:

$$\langle \Omega | b_{\uparrow}(\mathbf{k}) b_{\downarrow}(-\mathbf{k}) | \Omega \rangle = \Delta_{\mathbf{k}} \neq 0 \quad (16.44)$$

is the superconducting gap function: the expectation value of a two-particle operator. The same Fock-space mechanism applies: two fermions (individually forbidden from condensing by the Pauli exclusion principle) bind into a composite boson (which can condense). The gap $\Delta_{\mathbf{k}}$ is a Level 4 Born rendering of the pair field operator, with relative fluctuations suppressed by $1/\sqrt{N}$.

Preview: the Higgs as a relativistic condensate

The Higgs mechanism is the relativistic generalisation of Bose–Einstein condensation. The Higgs field $\hat{\phi}(x)$ has potential $V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4$, minimised at $|\phi| = v = \mu/\sqrt{2\lambda}$. The vacuum expectation value:

$$\langle \Omega | \hat{\phi}(x) | \Omega \rangle = v \quad (16.45)$$

is the relativistic order parameter: a Level 4 Born rendering of the Higgs field operator, deterministic in the infinite-volume limit. The $W^{\pm}$ and Z^0 acquire masses from the Higgs condensate exactly as Bogoliubov phonons acquire a gap from the BEC condensate; the eaten Goldstone bosons are the relativistic analogues of the superfluid phonons.

16.13 The Born rendering layer: consistency across levels

The degenerate Born rule applies at every scale. What changes is the degeneracy of the observable: finer observables have smaller eigensubspaces (less degeneracy); coarser observables have larger eigensubspaces (more degeneracy). The consistency of the hierarchy reduces to one fact: the

squared length of a bigger projection is the OR sum of the squared lengths of the smaller projections it contains.

The Born-Bohr layer must be postulated

The Schrödinger equation $\dot{h}_t = -iHh_t$ is linear, unitary, and deterministic. To extract a classical observation, three elements must be postulated:

1. The observable $\hat{O}$: eigenvalues define the outcomes; degeneracy determines the character of the output.
2. The Born probability: $P(\lambda) = \|P_\lambda h_t\|^2$ (squared length of the projected superposition).
3. The Bohr collapse: $h_{t+} = P_\lambda h_t / \sqrt{P(\lambda)}$ (trim and renormalize).

None follows from the Schrödinger equation. The Born-Bohr layer is a separate postulate. The question is whether different levels of coarse-graining give consistent results.

The four levels

Let the full Fock space factor as $\mathcal{F} = \mathcal{F}_S \otimes \mathcal{F}_E$ (system + environment), with hidden state $h_t = \sum_{a,e} c(a,e) |a,e\rangle$.

Level 1 ($d = 1$): single basis vector. Observable: complete measurement of both S and E . Projected superposition: $h_t^{(a_o, e_o)} = c(a_o, e_o) |a_o, e_o\rangle$. Born probability: $P(a_o, e_o) = |c(a_o, e_o)|^2$. Collapse: $h_{t+} = |a_o, e_o\rangle$. No OR sum. Quantum random.

Level 2 ($d = n_E$): system only. Observable: $\hat{O}_S \otimes \mathbf{1}_E$. Projected superposition: $h_t^{(a_o)} = \sum_e c(a_o, e) |a_o, e\rangle$. Born probability:

$$P(a_o) = \sum_e |c(a_o, e)|^2 = \sum_e P(a_o, e). \quad (16.46)$$

The OR sum: the Level 2 probability is the marginalisation of the Level 1 probability over e . The Pythagorean theorem guarantees consistency.

Level 3 ($d \sim e^S$): thermal. When E is a large bath, the ETH gives $P(a_o) = e^{-\beta E_{a_o}}/Z$: the Boltzmann distribution. The OR sum over $\sim e^{10^{23}}$ environment states, weighted by their amplitudes, produces the thermal probabilities. Variance: $\sim 1/\sqrt{N}$.

Level 4 ($d \sim e^{NS}$): macroscopic. Observable: collective $\hat{\Phi} = N^{-1} \sum_k \hat{\phi}_k$. The eigensubspace of Φ_o contains $\sim e^{NS(\Phi_o)}$ Fock configurations. Projected superposition:

$$h_t^{(\Phi_o)} = \sum_{\{a_k\}: \bar{\phi}=\Phi_o} c(\{a_k\}) |\{a_k\}\rangle. \quad (16.47)$$

Variance: $\text{Var}(\hat{\Phi}) \sim 1/N \rightarrow 0$. Born probability concentrates: $P(\hat{\Phi} \approx \langle \hat{\Phi} \rangle) \approx 1$. The Born output is deterministic.

The consistency chain

Each level's projected superposition is a union of the previous level's projected superpositions. The OR sum (Pythagorean theorem) guarantees:

$$\underbrace{|c(a_o, e_o)|^2}_{\substack{\text{Level 1} \\ d=1}} \xrightarrow{\sum_e} \underbrace{\sum_e |c(a_o, e)|^2}_{\substack{\text{Level 2} \\ d=n_E}} \xrightarrow{\text{ETH}} \underbrace{e^{-\beta E_{a_o}}/Z}_{\substack{\text{Level 3} \\ d \sim e^S}} \xrightarrow{N \rightarrow \infty} \underbrace{P(\Phi_o) \approx \delta(\Phi_o - \bar{\Phi})}_{\substack{\text{Level 4} \\ d \sim e^{NS}}} \quad (16.48)$$

Each arrow increases the degeneracy, decreases the variance, and makes the Born output more classical. No level contradicts any other: each coarser projection contains finer projections, and $\|h_t^{(\text{coarse})}\|^2 = \sum_{\text{fine}} \|h_t^{(\text{fine})}\|^2$.

Level	Degeneracy	Projected superposition	Variance	Born output
1	1	Single $ a_o, e_o\rangle$	Maximal	Quantum random
2	n_E	$\sum_e c(a_o, e) a_o, e\rangle$	$O(1)$	Marginal probability
3	$\sim e^{S/k_B}$	Same, with ETH	$\sim 1/\sqrt{N}$	Boltzmann
4	$\sim e^{NS}$	$\sum_{\{a_k\}} \dots$	$\sim 1/N \rightarrow 0$	Deterministic

There is no quantum-to-classical transition

The chain (16.48) makes clear: there is no moment at which h_t becomes classical. The hidden state is always a quantum superposition. The Born output is always a probability distribution. What changes is the *degeneracy*—the size of the projected superposition—which determines the variance and entropy of the Born distribution.

- **Small projected superposition** (Level 1, $d = 1$): one term, variance $O(1)$, entropy $O(\log \dim \mathcal{H})$. The Born output is quantum random.
- **Enormous projected superposition** (Level 4, $d \sim e^{NS}$): e^{NS} terms, variance $\sim 1/N \rightarrow 0$, entropy ~ 0 . The Born output is deterministic: classical.

The “quantumness” of a physical situation is the entropy of the Born output—determined by the degeneracy of the observable—not an intrinsic property of the hidden state. The Fock hidden state is always quantum. The classical world is what the Born distribution looks like when the projected superposition is so large that the law of large numbers makes the outcome certain.

16.14 Time slicing and the path integral in Fock language

The transition amplitude from initial Fock state $|i\rangle$ to final state $|f\rangle$ is the matrix element:

$$\mathcal{A}(i \rightarrow f; T) = \langle f | e^{-iHT} | i \rangle. \quad (16.49)$$

Time slicing: split $e^{-iHT} = (e^{-iH\varepsilon})^N$ with $\varepsilon = T/N$, and insert Fock resolutions of the identity $\mathbf{1} = \sum_n \int \prod_j \frac{d^3 \mathbf{k}_j}{(2\pi)^3} |n; \mathbf{k}_1, \dots\rangle \langle n; \mathbf{k}_1, \dots|$ between each factor:

$$\mathcal{A}(i \rightarrow f; T) = \sum_{\text{Fock configs}} \prod_{\text{steps}} \langle \text{config}_{j+1} | e^{-iH\varepsilon} | \text{config}_j \rangle. \quad (16.50)$$

The path integral is not a sum over classical field trajectories. This is the crucial point. The intermediate configurations in (16.50) are Fock basis vectors — specific particle configurations, one per time slice. A “path” is a sequence of Fock basis vectors: a sequence of particle configurations labelled by matrix multiplication indices. It is not a trajectory of a classical field $\psi(\mathbf{x}, t)$. It is a term in a matrix product.

When the Hamiltonian has the kinetic-plus-potential form, the short-time matrix element $\langle \mathbf{x}_{j+1} | e^{-iH\varepsilon} | \mathbf{x}_j \rangle$ can be evaluated by a Gaussian integral over the conjugate field (the momentum π), which performs the Legendre transform and produces the Lagrangian in the exponent. This is how the path integral $\int \mathcal{D}\psi e^{iS[\psi]}$ arises: it is a rewriting of the Fock matrix product in terms of field configurations, after the Legendre transform. The Lagrangian $\mathcal{L}$ is a derived object; the

Hamiltonian and the Fock space are the starting point. For condensed matter Hamiltonians without a simple kinetic-plus-potential form, one stays in the Hamiltonian language entirely, using the Dyson series instead.

16.15 The Dyson series and Feynman diagrams

Split $H = H_{\text{free}} + H_{\text{int}}$ and work in the interaction picture. The Dyson series:

$$U_I(T) = T \exp\left(-i \int_0^T H_{\text{int}}(t) dt\right) = \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int_0^T dt_1 \cdots dt_n T[H_{\text{int}}(t_1) \cdots H_{\text{int}}(t_n)] \quad (16.51)$$

generates the perturbative expansion of $\mathcal{A}(i \rightarrow f; T)$ in powers of H_{int} . Each term is a specific sequence of elementary interactions: n applications of H_{int} , each moving the Fock hidden state from one configuration to another via intermediate Fock states. Inserting resolutions of the identity between factors gives:

$$\mathcal{A}^{(2)}(i \rightarrow f; T) = (-i)^2 \sum_m \int_0^T dt_1 \int_0^{t_1} dt_2 \langle f | H_{\text{int}}(t_1) | m \rangle \langle m | H_{\text{int}}(t_2) | i \rangle. \quad (16.52)$$

The intermediate Fock state $|m\rangle$ is the *virtual state*: a summation index in the matrix product, not a physical state that is ever observed. Virtual states are not particles; they are terms in a computation.

Feynman diagrams. Each term in (16.51) is a Feynman diagram: external lines are initial and final Fock states; vertices are applications of H_{int} ; internal lines are free propagators $e^{-i\varepsilon_{\mathbf{k}}\tau}$ of virtual states between two vertices. The full amplitude is the sum of all diagrams at all orders. For the electron gas, the first-order diagram for electron-electron scattering is: two electrons exchange momentum $\mathbf{q}$ via $\tilde{V}(\mathbf{q}) = 4\pi e^2/\mathbf{q}^2$, the Coulomb vertex. The “virtual photon” internal line is the Coulomb propagator: a summation index, not a physical photon.

16.16 Preview: towards relativistic quantum field theory

The non-relativistic theory has one fundamental limitation: it is not Lorentz invariant. The dispersion relation $\varepsilon_{\mathbf{k}} = \mathbf{k}^2/2m$ is non-relativistic; the field operator (16.12) does not transform equivariantly under Lorentz boosts. The relativistic theory corrects this by adding structure to the basis vectors.

Lorentz equivariance requires spinor indices. A relativistic field must transform equivariantly: under a Lorentz transformation Λ , a spinor field transforms as $\psi_A(x) \rightarrow S(\Lambda)^A_B \psi_B(\Lambda^{-1}x)$, where $S(\Lambda)$ is the spinor representation matrix and A is the spinor index. This index must be added to the Fock basis vectors: $|\mathbf{k}, s\rangle$ carries spin s in the appropriate Lorentz representation. Without it, the theory cannot be consistent under Lorentz transformations.

Starting from a Lorentz-invariant Lagrangian. The cleanest route to a relativistic theory: write a Lagrangian $\mathcal{L}[\phi, \partial_\mu \phi]$ that is a Lorentz scalar; derive the Hamiltonian H by Legendre transform; build the Fock space with basis vectors $|\mathbf{k}, s, \dots\rangle$ carrying the appropriate spinor and internal indices; and time-slice e^{-iHT} to recover the path integral $\int \mathcal{D}\phi e^{iS}$. The round trip (??):

$$\mathcal{L} \rightarrow H \rightarrow \mathcal{F} \rightarrow e^{-iHT} \rightarrow \int \mathcal{D}\phi e^{iS} \rightarrow \mathcal{L}$$

is the standard procedure of relativistic QFT. For fermions, the intermediate field variables at each time slice are Grassmann numbers — anticommuting classical fields — not ordinary commuting numbers.

Adding internal indices: gauge coupling. The Standard Model further adds colour $\alpha \in \{r, g, b\}$, isospin $i \in \{1, 2\}$, and hypercharge to the Fock basis vectors. These indices transform under the gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$ by internal matrices. Requiring the Lagrangian to be gauge-invariant forces the introduction of gauge fields (the photon, $W^\pm$, Z^0 , gluons), which are themselves Fock degrees of freedom. All forces arise from this single requirement.

The bridge, completed. The non-relativistic Fock-Born theory of this section is the Standard Model with simpler basis vectors: no Lorentz spinor index, no colour or isospin, non-relativistic dispersion $\varepsilon_{\mathbf{k}} = \mathbf{k}^2/2m$. The bridge to the Standard Model is:

Non-relativistic QFT	Standard Model
Basis: $ n; \mathbf{k}_j, s_j\rangle$	Basis: $ n; \mathbf{k}_j, s_j, \alpha_j, i_j, f_j, \pm_j\rangle$
Dispersion: $\mathbf{k}^2/2m$	Dispersion: $\sqrt{\mathbf{k}^2 + m^2}$
No spinor structure	Lorentz spinor index $A \in \{1, 2, 3, 4\}$
No gauge index	Colour α , isospin i , hypercharge Y
Coulomb interaction	Gauge coupling (QED, QCD, weak)
H is 2-3 parameters	H_{SM} is 19 parameters
Path integral: $\int \mathcal{D}\psi e^{iS}$	Same, with Grassmann fields for fermions
Born: density, spectral function	Born: cross sections, decay rates, condensates

The Schrödinger equation $\dot{h}_t = -iHh_t$, the Fock space $\mathcal{F} = \bigoplus_n \mathcal{H}^{(n)}$, the Born rendering layer, and the Dyson series are unchanged. The Standard Model is non-relativistic QFT with richer basis vectors and a structured Hamiltonian.

17 Lorentz Transformations, Rotations, Boosts, and Einstein notation

In the non-relativistic quantum field theory of the previous section, time and space played fundamentally different roles. Time was the external parameter of the Schrödinger equation $\dot{h}_t = -iHh_t$: the game engine ticks forward in t , and the Fock hidden state h_t is a spatial snapshot at each tick. The spatial momentum $\mathbf{k} \in \mathbb{R}^3$ labelled the Fock basis vectors; the energy $\varepsilon_{\mathbf{k}} = \mathbf{k}^2/2m$ was a function of spatial momentum alone. Time and space entered the theory on completely different footings, and there was no requirement that the physics look the same to observers moving at different velocities.

Special relativity removes this asymmetry. Time and space are unified into a single four-dimensional spacetime, with coordinates $x^\mu = (t, \mathbf{x})$, and the fundamental symmetry group of the theory is the *Lorentz group*: the group of linear transformations Λ of spacetime that preserve the Minkowski interval $s^2 = t^2 - |\mathbf{x}|^2$. A Lorentz transformation mixes time and space, exactly as an ordinary rotation mixes the spatial coordinates x , y , and z . The difference is the sign in the metric: while a rotation preserves $x^2 + y^2 + z^2$, a Lorentz boost preserves $t^2 - x^2 - y^2 - z^2$. This sign — the difference between Euclidean and Minkowski geometry — is the origin of the factor of i in the Wick rotation $t \rightarrow -i\tau$: rotating to imaginary time converts the Minkowski metric into the Euclidean metric $\tau^2 + |\mathbf{x}|^2$, turning Lorentz transformations into ordinary rotations in four-dimensional Euclidean space. The path integral in Euclidean time, which we used to project onto the ground state and compute instanton amplitudes, is the analytic continuation of the Minkowski path integral through this Wick rotation.

For a quantum field theory to be consistent with special relativity, the requirement is surprisingly simple to state: the Lagrangian density $\mathcal{L}$ must be a *Lorentz scalar*. That is, under any Lorentz transformation Λ , the Lagrangian density at the transformed spacetime point must equal the original: $\mathcal{L}[\phi(\Lambda x)] = \mathcal{L}[\phi(x)]$. The action $S = \int d^4x \mathcal{L}$ is then Lorentz invariant — the four-volume element d^4x is also preserved by Lorentz transformations — and the equations of motion derived from it are automatically Lorentz covariant. The individual fields $\phi(x)$, $A^\mu(x)$, $\psi_A(x)$ appearing in $\mathcal{L}$ may transform non-trivially under Λ , carrying Lorentz indices such as the vector index μ or the spinor index A . But these indices must be contracted in the Lagrangian in a way that makes the whole expression a scalar: every free index must be summed over with the appropriate metric or invariant tensor. This single requirement — Lorentz scalar Lagrangian — determines which combinations of fields are allowed as interaction terms, forces the Dirac equation on spinor fields, and constrains the form of every vertex in the Standard Model. To implement it correctly, we must understand how different fields transform under Lorentz transformations and how to build scalars from them: this is the representation theory of the Lorentz group, which we take up in the next section.

17.1 Spacetime and the invariant interval

In special relativity, spacetime is $\mathbb{R}^4$ with coordinates $x^\mu = (x^0, x^1, x^2, x^3) = (ct, x, y, z)$ (setting $c = 1$ throughout). The fundamental invariant is the *spacetime interval*:

$$s^2 = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 = \eta_{\mu\nu} x^\mu x^\nu, \quad (17.1)$$

where $\eta_{\mu\nu} = \text{diag}(+1, -1, -1, -1)$ is the *Minkowski metric*. The interval s^2 is the same in every inertial frame. Positive s^2 is timelike (causally connected), zero is lightlike (on the light cone), negative is spacelike (causally disconnected).

ML reading. The Minkowski metric $\eta_{\mu\nu}$ is the inner product matrix of spacetime. Ordinary Euclidean space has inner product matrix $\delta_{\mu\nu} = \text{diag}(+1, +1, +1)$; Minkowski space has $\eta_{\mu\nu} =$

diag(+1, -1, -1, -1). The sign flip on the spatial components is what distinguishes time from space, and it is the source of all the counterintuitive features of special relativity.

17.2 Einstein index notation: upper and lower indices

Physics uses a compact notation that makes Lorentz invariance manifest. Every repeated index—one upper, one lower—implies a sum. This is the *Einstein summation convention*:

$$\eta_{\mu\nu} x^\mu x^\nu \equiv \sum_{\mu=0}^3 \sum_{\nu=0}^3 \eta_{\mu\nu} x^\mu x^\nu. \quad (17.2)$$

The rule is: a free index (appearing once) labels a component; a contracted index (appearing twice, once up and once down) is summed over.

Upper indices (contravariant components). x^μ is a *contravariant* vector—it transforms by the Lorentz matrix Λ :

$$x^\mu \mapsto x'^\mu = \Lambda^\mu{}_\nu x^\nu. \quad (17.3)$$

Here $\Lambda^\mu{}_\nu$ is the Lorentz matrix with the first index (row) up and the second index (column) down.

Lower indices (covariant components). The metric $\eta_{\mu\nu}$ is used to *lower* an index:

$$x_\mu = \eta_{\mu\nu} x^\nu = (x^0, -x^1, -x^2, -x^3). \quad (17.4)$$

The covariant vector x_μ transforms by the inverse-transpose of Λ :

$$x_\mu \mapsto x'_\mu = (\Lambda^{-T})_\mu{}^\nu x_\nu = \Lambda_\mu{}^\nu x_\nu. \quad (17.5)$$

A contracted upper-lower pair $x^\mu x_\mu = x^\mu \eta_{\mu\nu} x^\nu = s^2$ is Lorentz invariant: it has no free indices.

The inverse metric. The inverse of $\eta_{\mu\nu}$ is $\eta^{\mu\nu} = \text{diag}(+1, -1, -1, -1)$ (coincidentally the same matrix for Minkowski), used to *raise* an index:

$$x^\mu = \eta^{\mu\nu} x_\nu. \quad (17.6)$$

Derivatives. The partial derivative $\partial/\partial x^\mu$ with respect to a contravariant coordinate is itself *covariant*—it carries a lower index:

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right). \quad (17.7)$$

The d'Alembertian (wave operator) is the Lorentz-invariant contraction:

$$\partial^\mu \partial_\mu = \eta^{\mu\nu} \partial_\mu \partial_\nu = \partial_t^2 - \nabla^2 = \square. \quad (17.8)$$

The Klein–Gordon equation $(\square + m^2)\phi = 0$ is Lorentz invariant because $\square$ is a contracted pair with no free indices.

Example. Index counting in $\mathcal{L} = \frac{1}{2}(\partial_\mu \phi)(\partial^\mu \phi) - \frac{1}{2}m^2 \phi^2$.

The kinetic term: $\partial_\mu \phi$ has one lower free index μ ; $\partial^\mu \phi = \eta^{\mu\nu} \partial_\nu \phi$ has one upper free index μ . Contracting: $(\partial_\mu \phi)(\partial^\mu \phi)$ has no free indices—it is a Lorentz scalar. The mass term $m^2 \phi^2$ has no indices. Every term in $\mathcal{L}$ is a scalar. Since d^4x is also a scalar (the Lorentz Jacobian $|\det \Lambda| = 1$), the action $S = \int d^4x \mathcal{L}$ is invariant. ✓

17.3 Lorentz transformations: definition

A *Lorentz transformation* is a 4×4 real matrix $\Lambda^\mu{}_\nu$ that preserves the spacetime interval (17.1):

$$\eta_{\mu\nu} \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma = \eta_{\rho\sigma}, \quad \text{i.e.} \quad \Lambda^T \eta \Lambda = \eta. \quad (17.9)$$

In matrix notation: $\Lambda^T \eta \Lambda = \eta$. This is the defining equation of the *Lorentz group* $O(3,1)$. Taking the determinant: $(\det \Lambda)^2 = 1$, so $\det \Lambda = \pm 1$.

The *proper orthochronous* Lorentz group $SO^+(3,1)$ consists of transformations with $\det \Lambda = +1$ and $\Lambda^0{}_0 \geq 1$ (time-orientation preserving). This is the physically relevant subgroup for continuous transformations connected to the identity. It has six parameters: three rotations and three boosts.

17.4 Rotations

A *rotation* is a Lorentz transformation that acts only on the spatial components and leaves $x^0 = t$ unchanged. For a rotation by angle θ about the z -axis:

$$\Lambda_{\text{rot},z}(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta & 0 \\ 0 & \sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (17.10)$$

Rotations form a subgroup $SO(3) \subset SO^+(3,1)$. They preserve both the time component x^0 and the spatial length $|\mathbf{x}|^2$ separately, as well as the full interval s^2 .

The generators of infinitesimal rotations are the angular momentum matrices J_i . For rotation about the z -axis:

$$J_z = -i \frac{d}{d\theta} \Lambda_{\text{rot},z} \Big|_{\theta=0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (17.11)$$

so $\Lambda_{\text{rot},z}(\theta) = e^{-i\theta J_z}$. The three rotation generators satisfy $[J_i, J_j] = i\epsilon_{ijk} J_k$ —the $\mathfrak{su}(2) \cong \mathfrak{so}(3)$ algebra, which is the same algebra as the Pauli spin matrices (but now acting on spacetime rather than on spinors).

17.5 Boosts

A *boost* is a Lorentz transformation that mixes the time component x^0 with one spatial component. For a boost with velocity v along the x -axis (setting $c = 1$):

$$\Lambda_{\text{boost},x}(v) = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \gamma = \frac{1}{\sqrt{1-v^2}}. \quad (17.12)$$

It is conventional and algebraically cleaner to parametrise by the *rapidity* $\xi = \tanh^{-1} v$, so that $\gamma = \cosh \xi$ and $\gamma v = \sinh \xi$:

$$\Lambda_{\text{boost},x}(\xi) = \begin{pmatrix} \cosh \xi & -\sinh \xi & 0 & 0 \\ -\sinh \xi & \cosh \xi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = e^{-i\xi K_x}, \quad (17.13)$$

where K_x is the boost generator along x :

$$K_x = -i \frac{d}{d\xi} \Lambda_{\text{boost},x} \Big|_{\xi=0} = \begin{pmatrix} 0 & -i & 0 & 0 \\ -i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (17.14)$$

Rapidity adds linearly for collinear boosts: two boosts of rapidity ξ_1 and ξ_2 along the same axis compose to rapidity $\xi_1 + \xi_2$. This is the Lorentz analogue of angle addition for rotations.

Key difference from rotations. Rotations involve sin and cos; boosts involve sinh and cosh. The rotation matrix is orthogonal ($\Lambda^T \Lambda = I$); the boost matrix is not orthogonal in the Euclidean sense, but satisfies the Lorentz condition $\Lambda^T \eta \Lambda = \eta$. Rotations act on a compact parameter space ($\theta \in [0, 2\pi)$); boosts act on a non-compact parameter space ($\xi \in (-\infty, \infty)$). This non-compactness means that finite-dimensional representations of the Lorentz group cannot be unitary—a fact that forces spinors to transform non-unitarily under boosts.

17.6 The Lie algebra of the Lorentz group

The full Lorentz group $SO^+(3, 1)$ has six generators: three rotations J_i and three boosts K_i ($i = x, y, z$). Their commutation relations are:

$$[J_i, J_j] = i \epsilon_{ijk} J_k, \quad (17.15)$$

$$[K_i, K_j] = -i \epsilon_{ijk} J_k, \quad (17.16)$$

$$[J_i, K_j] = i \epsilon_{ijk} K_k. \quad (17.17)$$

The first relation (17.15) is the $\mathfrak{su}(2)$ rotation algebra. The second (17.16) says that two boosts in different directions produce a rotation (the *Wigner rotation*—the physical origin of Thomas precession). The third (17.17) says that boosts transform as vectors under rotations.

Here ϵ_{ijk} is the Levi-Civita symbol, totally antisymmetric in all three indices with $\epsilon_{xyz} = +1$, so that $\epsilon_{ijk} = +1$ for even permutations of (x, y, z) , $\epsilon_{ijk} = -1$ for odd permutations, and $\epsilon_{ijk} = 0$ if any two indices coincide; it is the standard way to write the cross product in index notation, encoding the right-hand rule of three-dimensional rotations.

The algebra (17.15)–(17.17) is $\mathfrak{so}(3, 1)$. Defining the complexified combinations:

$$A_i = \frac{1}{2}(J_i + iK_i), \quad B_i = \frac{1}{2}(J_i - iK_i), \quad (17.18)$$

one finds $[A_i, A_j] = i \epsilon_{ijk} A_k$, $[B_i, B_j] = i \epsilon_{ijk} B_k$, $[A_i, B_j] = 0$. The Lorentz algebra *factorises* into two independent $\mathfrak{su}(2)$ algebras. This is the algebraic origin of the (j_L, j_R) classification of representations introduced in diagram 3.

Example. Lorentz invariant from index contraction.

The electromagnetic field-strength tensor $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ has two lower indices. Under a Lorentz transformation:

$$F_{\mu\nu} \mapsto F'_{\mu\nu} = \Lambda_\mu{}^\rho \Lambda_\nu{}^\sigma F_{\rho\sigma}.$$

To form a scalar, raise both indices with $\eta^{\mu\rho} \eta^{\nu\sigma}$ to get $F^{\mu\nu}$, then contract:

$$F^{\mu\nu} F_{\mu\nu} = \eta^{\mu\rho} \eta^{\nu\sigma} F_{\rho\sigma} F_{\mu\nu}.$$

All four indices are contracted: no free indices remain, so this is a Lorentz scalar. Evaluating explicitly: $F^{\mu\nu} F_{\mu\nu} = 2(|\mathbf{B}|^2 - |\mathbf{E}|^2)$. The Maxwell Lagrangian $\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu}$ is therefore Lorentz invariant by index counting alone.

17.7 Lorentz invariance: the practical test

An expression is Lorentz invariant if and only if it has *no free Lorentz indices*. The practical test:

Expression	Free indices	Lorentz status
$x^\mu x_\mu = \eta_{\mu\nu} x^\mu x^\nu$	none	Scalar (invariant)
$\partial_\mu \phi \partial^\mu \phi$	none	Scalar (invariant)
$F_{\mu\nu} F^{\mu\nu}$	none	Scalar (invariant)
x^μ	μ	4-vector (covariant)
$F_{\mu\nu}$	μ, ν	rank-2 tensor
$\partial_\mu F^{\mu\nu}$	ν	4-vector (Maxwell eqn lhs)
$g_{\mu\nu} R^{\mu\nu} = R$	none	Scalar (Ricci scalar)

The Lagrangian density $\mathcal{L}$ must be a scalar (no free indices) for the action $S = \int d^4x \mathcal{L}$ to be Lorentz invariant. The spacetime measure d^4x transforms with $|\det \Lambda| = 1$, so it contributes no index structure. Writing a Lorentz-invariant Lagrangian is therefore the same as writing an expression with all indices contracted.

Summary: upper vs. lower at a glance.

Object	Index position	Transforms as
Position $x^\mu = (t, \mathbf{x})$	upper	$x'^\mu = \Lambda^\mu{}_\nu x^\nu$
Gradient $\partial_\mu = (\partial_t, \nabla)$	lower	$\partial'_\mu = \Lambda_\mu{}^\nu \partial_\nu$
Metric $\eta_{\mu\nu}$	two lower	invariant tensor
Inverse metric $\eta^{\mu\nu}$	two upper	invariant tensor
Field A_μ (photon potential)	lower	$A'_\mu = \Lambda_\mu{}^\nu A_\nu$
Field strength $F^{\mu\nu}$	two upper	$F'^{\mu\nu} = \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma F^{\rho\sigma}$
Contracted pair $V^\mu W_\mu$	none	Lorentz scalar

17.8 Wick rotation: Lorentz as a rotated sphere

Wick rotation is one of the most illuminating ways to understand the Lorentz group, and it also turns out to be the key technical step that connects the Minkowski path integral to the Euclidean path integral used in lattice QCD and statistical mechanics.

The Minkowski interval and its sign problem

The spacetime interval in Minkowski space is:

$$s^2 = t^2 - x^2 - y^2 - z^2 = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2. \quad (17.19)$$

The mixed signs make this look different from an ordinary Euclidean distance. In four-dimensional Euclidean space the distance squared is:

$$r^2 = x_E^2 + y_E^2 + z_E^2 + \tau^2, \quad (17.20)$$

which has all positive signs and is invariant under $SO(4)$ —the group of rotations in four dimensions. The Lorentz group $SO(1, 3)$ (or more precisely $SO^+(3, 1)$) is the symmetry group of (17.19). The two groups have the same number of generators (six) but different algebraic structures.

The question is: can the mixed signs in (17.19) be transformed into all-positive signs by a change of variable?

Wick rotation: $t \rightarrow -i\tau$

Define the *Euclidean time* τ by rotating the time coordinate into the imaginary axis:

$$t = x^0 \mapsto -i\tau, \quad \tau \in \mathbb{R}. \quad (17.21)$$

Then:

$$t^2 = (-i\tau)^2 = -\tau^2, \quad (17.22)$$

and the Minkowski interval becomes:

$$s^2 = t^2 - x^2 - y^2 - z^2 = -\tau^2 - x^2 - y^2 - z^2 = -(r_E^2), \quad (17.23)$$

where $r_E^2 = \tau^2 + x^2 + y^2 + z^2$ is the Euclidean distance squared with all positive signs. Up to the overall minus sign (a convention), the Wick-rotated interval is exactly the four-dimensional Euclidean distance.

The symmetry group of r_E^2 is $SO(4)$ —the group of rotations of the four-dimensional sphere S^3 . So after Wick rotation, the symmetry group of spacetime becomes $SO(4)$ instead of $SO^+(3,1)$. Lorentz boosts, which in Minkowski space mix t and x via hyperbolic functions ($\cosh \xi$, $\sinh \xi$), become ordinary rotations in the (τ, x) -plane via circular functions ($\cos \theta$, $\sin \theta$) after Wick rotation.

Boosts become rotations

In Minkowski space, a boost with rapidity ξ along the x -axis is:

$$\begin{pmatrix} t' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh \xi & \sinh \xi \\ \sinh \xi & \cosh \xi \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix}. \quad (17.24)$$

Substitute $t = -i\tau$, $t' = -i\tau'$:

$$\begin{pmatrix} -i\tau' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh \xi & \sinh \xi \\ \sinh \xi & \cosh \xi \end{pmatrix} \begin{pmatrix} -i\tau \\ x \end{pmatrix}. \quad (17.25)$$

This gives $\tau' = \tau \cosh \xi - ix \sinh \xi$ and $x' = -i\tau \sinh \xi + x \cosh \xi$. For this to be a real rotation of real (τ, x) , we need the boost parameter to be imaginary: $\xi \rightarrow i\theta$. Then $\cosh(i\theta) = \cos \theta$ and $\sinh(i\theta) = i \sin \theta$, giving:

$$\begin{pmatrix} \tau' \\ x' \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \tau \\ x \end{pmatrix}. \quad (17.26)$$

This is an ordinary $SO(2)$ rotation in the (τ, x) -plane. The hyperbolic boost has become a circular rotation.

Example. The relationship between $SO^+(3,1)$ and $SO(4)$.

The Lie algebras of $SO^+(3,1)$ and $SO(4)$ are both six-dimensional. The rotation generators J_i are the same in both groups. The difference is in the boost generators K_i : in $SO^+(3,1)$ they are Hermitian (boosts are non-compact), while in $SO(4)$ the corresponding generators $\tilde{K}_i = iK_i$ are anti-Hermitian (rotations are compact). Wick rotation is exactly the substitution $K_i \rightarrow \tilde{K}_i = iK_i$: it turns the non-compact boost generators into compact rotation generators.

The $A_i = \frac{1}{2}(J_i + iK_i)$ and $B_i = \frac{1}{2}(J_i - iK_i)$ decomposition is the complexification that makes this structure manifest: after Wick rotation, A_i and B_i both become $\mathfrak{su}(2)$ generators of $SO(4) \cong SU(2) \times SU(2)/\mathbb{Z}_2$.

The Euclidean path integral

Wick rotation is not just a geometrical curiosity—it is the key computational tool that converts the oscillatory Minkowski path integral into a convergent Euclidean one.

The Minkowski path integral has the weight $e^{iS_M[\phi]}$, which oscillates with modulus 1 and is hard to evaluate numerically. Under $t \rightarrow -i\tau$, the Minkowski action $S_M = \int dt \mathcal{L}_M$ becomes:

$$iS_M = i \int dt \mathcal{L}_M = i \int (-id\tau) \mathcal{L}_M = \int d\tau \mathcal{L}_M = -S_E, \quad (17.27)$$

where $S_E = \int d\tau (-\mathcal{L}_M)|_{t=-i\tau}$ is the *Euclidean action*, which is real and positive for stable theories. The path integral weight becomes:

$$e^{iS_M} \mapsto e^{-S_E}, \quad (17.28)$$

a real, positive, decaying exponential. The Euclidean path integral $\int \mathcal{D}\phi e^{-S_E[\phi]}$ is now a well-defined probability measure—exactly a Boltzmann distribution with S_E playing the role of energy divided by temperature. This is why the path integral of a QFT in Euclidean signature is mathematically equivalent to the partition function of a statistical field theory.

Concept	Minkowski	Euclidean (after Wick)
Time coordinate	$t \in \mathbb{R}$	$\tau = it \in \mathbb{R}$
Metric signature	$(+, -, -, -)$	$(+, +, +, +)$
Interval	$t^2 - \mathbf{x} ^2$	$\tau^2 + \mathbf{x} ^2$
Symmetry group	$SO^+(3, 1)$ (non-compact)	$SO(4)$ (compact)
Boosts	Hyperbolic (cosh, sinh)	Circular (cos, sin)
Path integral weight	e^{iS_M} (oscillatory)	e^{-S_E} (decaying)
Interpretation	Quantum amplitude	Statistical weight

The 4-sphere picture

The intuition is that $s^2 = t^2 - |\mathbf{x}|^2 = \text{const}$ looks like a sphere after Wick rotation. The hyperboloid $t^2 - |\mathbf{x}|^2 = R^2$ in Minkowski space (the set of all points at fixed proper time R from the origin) becomes the 3-sphere $\tau^2 + |\mathbf{x}|^2 = R^2$ after $t \rightarrow -i\tau$. Minkowski hyperboloids are the “Wick-rotated” version of Euclidean spheres.

The Lorentz group $SO^+(3, 1)$ is the symmetry group of the Minkowski hyperboloid. Its Euclidean counterpart $SO(4)$ is the symmetry group of the 3-sphere S^3 . The two groups have the same Lie algebra over the complex numbers (they are real forms of the same complex Lie algebra), but they differ in their global topology: $SO(4)$ is compact (finite volume) while $SO^+(3, 1)$ is non-compact (infinite volume, because rapidities are unbounded). This non-compactness is why finite-dimensional representations of $SO^+(3, 1)$ cannot be unitary—in a compact group, all finite-dimensional representations are unitary; in a non-compact group, they need not be.

ML reading. Wick rotation is a change of basis in the time direction from real to imaginary. In the RNN picture, the Schrödinger equation $\dot{h}_t = -iHh_t$ becomes under $t \rightarrow -i\tau$:

$$\frac{dh_\tau}{d\tau} = -H h_\tau. \quad (17.29)$$

The evolution operator e^{-iHt} (unitary, oscillating) becomes $e^{-H\tau}$ (real, decaying). This is the transition from a quantum amplitude to a classical probability: the Euclidean RNN is a

contraction mapping (for positive H) rather than a rotation. The ground state of H dominates as $\tau \rightarrow \infty$, which is why Euclidean methods are used to extract the ground-state energy (e.g., the proton mass in lattice QCD): the Euclidean evolution filters out all excited states, leaving only the lowest eigenvalue.

17.9 Tensors: the chain rule in disguise

A tensor is not a mysterious object. It is simply a quantity that transforms under a change of coordinates by the chain rule. Once this is understood, the entire hierarchy of scalars, vectors, covectors, and higher-rank tensors becomes a single idea applied at different levels.

Coordinates and the chain rule

Let $x^\mu = (x^0, x^1, x^2, x^3)$ be coordinates in one inertial frame and x'^ν coordinates in another, related by a Lorentz transformation:

$$x'^\nu = \Lambda^\nu_\mu x^\mu. \quad (17.30)$$

The chain rule tells you how partial derivatives transform. If f is any function of spacetime:

$$\frac{\partial f}{\partial x'^\nu} = \frac{\partial x^\mu}{\partial x'^\nu} \frac{\partial f}{\partial x^\mu} = (\Lambda^{-1})^\mu_\nu \frac{\partial f}{\partial x^\mu}. \quad (17.31)$$

The chain rule is the origin of all tensor transformation laws. Quantities that transform *like coordinates* x^μ are contravariant; quantities that transform *like derivatives* $\partial/\partial x^\mu$ are covariant. Everything else is built from these two building blocks.

Contravariant vectors: quantities that transform like coordinates

The coordinate 4-vector x^μ transforms with the matrix Λ^ν_μ . Any quantity V^μ transforming by the same rule:

$$V'^\nu = \Lambda^\nu_\mu V^\mu \quad (17.32)$$

is a *contravariant 4-vector*. The upper index signals contravariance.

The 4-velocity of a particle. A particle traces a worldline $x^\mu(\tau)$ parametrised by proper time τ . The 4-velocity:

$$u^\mu = \frac{dx^\mu}{d\tau} = (\gamma c, \gamma \mathbf{v}), \quad \gamma = \frac{1}{\sqrt{1 - v^2/c^2}}, \quad (17.33)$$

transforms as a contravariant vector because the chain rule gives:

$$u'^\nu = \frac{dx'^\nu}{d\tau} = \frac{\partial x'^\nu}{\partial x^\mu} \frac{dx^\mu}{d\tau} = \Lambda^\nu_\mu u^\mu. \quad (17.34)$$

The transformation law is just the chain rule applied along the worldline. Similarly, the 4-momentum $p^\mu = mu^\mu = (E/c, \mathbf{p})$ is a contravariant vector.

Covariant vectors: quantities that transform like derivatives

The gradient of a scalar transforms with (Λ^{-1}) :

$$\partial'_\nu f = (\Lambda^{-1})^\mu_\nu \partial_\mu f. \quad (17.35)$$

Any W_μ transforming by this rule is a *covariant 4-vector* (covector); lower index signals covariance. The metric $\eta_{\mu\nu} = \text{diag}(-1, +1, +1, +1)$ converts between the two: $V_\mu = \eta_{\mu\nu} V^\nu$.

The inner product is a scalar. Contracting a covariant with a contravariant vector:

$$W_\mu V^\mu \rightarrow (\Lambda^{-1})^\mu{}_\nu W_\mu \cdot \Lambda^\nu{}_\rho V^\rho = \delta^\mu{}_\rho W_\mu V^\rho = W_\mu V^\mu. \quad (17.36)$$

The two transformation matrices cancel. A contracted pair of upper-lower indices is always Lorentz-invariant. This is Einstein summation convention: a repeated upper-lower index is summed and yields a scalar.

Rank-2 tensors: two indices, two chain rules

A rank-(2,0) tensor transforms with two factors of Λ :

$$T'^{\mu\nu} = \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma T^{\rho\sigma}. \quad (17.37)$$

The pattern: each index gets one chain-rule factor—upper indices get Λ , lower indices get Λ^{-1} . A rank-(p, q) tensor transforms with p factors of Λ and q factors of Λ^{-1} . Contracting one upper with one lower index reduces rank by (1,1) and gives a Lorentz-invariant.

The electromagnetic field strength tensor

The electromagnetic field is an antisymmetric rank-(0,2) tensor:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = -F_{\nu\mu}, \quad (17.38)$$

where $A_\mu = (-\phi/c, \mathbf{A})$ is the 4-potential. Antisymmetry sets the diagonal to zero and leaves $\binom{4}{2} = 6$ independent components—exactly the three electric and three magnetic field components:

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & B_z & -B_y \\ E_y/c & -B_z & 0 & B_x \\ E_z/c & B_y & -B_x & 0 \end{pmatrix}. \quad (17.39)$$

Under a Lorentz transformation, $F'_{\mu\nu} = (\Lambda^{-1})^\rho{}_\mu (\Lambda^{-1})^\sigma{}_\nu F_{\rho\sigma}$: the electric and magnetic fields mix. For a boost with velocity v along x :

$$E'_x = E_x, \quad E'_y = \gamma(E_y - vB_z), \quad E'_z = \gamma(E_z + vB_y), \quad (17.40)$$

and analogously for $\mathbf{B}$. A purely electric field in one frame has both $\mathbf{E}$ and $\mathbf{B}$ in another. The fundamental object is $F_{\mu\nu}$; $\mathbf{E}$ and $\mathbf{B}$ are its frame-dependent projections.

Classical electromagnetism in tensor notation

In tensor notation, Maxwell's four equations become two:

$$\partial^\mu F_{\mu\nu} = \mu_0 J_\nu, \quad (17.41)$$

$$\partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} + \partial_\rho F_{\mu\nu} = 0. \quad (17.42)$$

Equation (17.41) is Gauss's law and Ampère's law; (17.42) is Faraday's law and no magnetic monopoles. The 4-current is $J^\mu = (c\rho, \mathbf{J})$. Both equations are manifestly Lorentz-covariant by construction.

The Lagrangian:

$$\mathcal{L}_{\text{EM}} = -\frac{1}{4\mu_0} F_{\mu\nu} F^{\mu\nu} = \frac{\epsilon_0}{2} E^2 - \frac{1}{2\mu_0} B^2. \quad (17.43)$$

The contraction $F_{\mu\nu} F^{\mu\nu}$ is Lorentz-invariant (all indices contracted) and equals $2(B^2 - E^2/c^2)$ up to constants. Its Euler–Lagrange equations give (17.41).

A second invariant built from $F_{\mu\nu}$ is the pseudoscalar $F_{\mu\nu}\tilde{F}^{\mu\nu} \propto \mathbf{E} \cdot \mathbf{B}$, where $\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}$ is the dual. This changes sign under parity and appears in anomaly calculations. The two invariants $F_{\mu\nu}F^{\mu\nu}$ and $F_{\mu\nu}\tilde{F}^{\mu\nu}$ are the only independent algebraic invariants of the EM field.

The kinetic term $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ is the template for all gauge boson kinetic terms in the Standard Model: for gluons, replace $F_{\mu\nu}$ by the non-Abelian $G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc}G_\mu^b G_\nu^c$ and sum over the adjoint colour index $a = 1, \dots, 8$.

Foreshadowing spinors: when the chain rule is not enough

Every tensor representation is built from the chain-rule matrix $\Lambda^\mu{}_\nu$: each index gets one factor. But the Lorentz group $SO^+(1,3)$ has representations that *cannot* be described by any number of chain-rule factors. These are the spinor representations, and they are the key to understanding the electron.

Why spinors are not tensors. We will show that $SL(2, \mathbb{C})$ is the double cover of $SO^+(1,3)$: the two elements L and $-L$ in $SL(2, \mathbb{C})$ give the same Lorentz transformation Λ . A tensor representation depends only on Λ , not on the sign of L : a 360° rotation sends every tensor back to itself.

A spinor ψ transforms directly under L : a 360° rotation sends $L \rightarrow -L$, so $\psi \rightarrow -\psi$. No tensor can do this. The spinor representation cannot be built from $\Lambda^\mu{}_\nu$ by any finite number of index contractions—it is genuinely outside the tensor hierarchy.

The two fundamental spinors. $SL(2, \mathbb{C})$ has two inequivalent 2-dimensional representations:

- **Left Weyl spinor** ψ_L^m , $m \in \{1, 2\}$: transforms as $\psi_L \rightarrow L\psi_L$. Undotted index.
- **Right Weyl spinor** $\psi_R^{\dot{m}}$, $\dot{m} \in \{\dot{1}, \dot{2}\}$: transforms as $\psi_R \rightarrow (L^\dagger)^{-1}\psi_R$. Dotted index.

The 4-vector sits in the bi-spinor representation $(\frac{1}{2}, \frac{1}{2})$: it can be written as a 2×2 matrix $V^{m\dot{m}}$ via $V^\mu\sigma_\mu$, transforming as $V \rightarrow LV L^\dagger$ —one spinor index of each type.

The full classification.

(j_L, j_R)	Dim	Name	Indices	Example
$(0, 0)$	1	Scalar	none	ϕ , Higgs
$(\frac{1}{2}, 0)$	2	Left Weyl spinor	m	ψ_L , ν_L
$(0, \frac{1}{2})$	2	Right Weyl spinor	$\dot{m}$	ψ_R , e_R
$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	4	Dirac spinor	$m, \dot{m}$	$\Psi = \psi_L \oplus \psi_R$
$(\frac{1}{2}, \frac{1}{2})$	4	4-vector	μ	A^μ , p^μ
$(1, 0) \oplus (0, 1)$	6	Antisymm. tensor	$[\mu\nu]$	$F_{\mu\nu}$

Integer $j_L + j_R$ gives tensor (bosonic) representations; half-integer gives spinor (fermionic) representations. The spin-statistics theorem—which we foreshadow here and develop later—says that this is not a coincidence: Lorentz invariance and causality force integer-spin fields to be bosons and half-integer-spin fields to be fermions. The chain rule captures integer spin; the double cover captures half-integer spin. Together they classify every field in the Standard Model.

17.10 Lorentz invariance from the lattice

The Fock-Born framework defines the Fock layer on a discrete lattice with spacing a and a nearest-neighbour Hamiltonian $H(a)$. The Minkowski metric $\eta_{\mu\nu}$ has so far been treated as a given. We now show that $\eta_{\mu\nu}$ —and with it the full Lorentz invariance $SO^+(3,1)$ —emerges from

two properties of the lattice that we have already established: nearest-neighbour locality and stable Born rendering. No additional postulate is needed.

Ingredient 1: nearest-neighbour locality gives a light cone

Consider a Hamiltonian H on a lattice of spacing a , with interactions restricted to nearest neighbours. The Lieb–Robinson theorem (1972) proves rigorously that for any such Hamiltonian, the commutator of two local operators satisfies:

$$||[\hat{A}(\mathbf{x}, 0), \hat{B}(\mathbf{y}, t)]|| \leq C e^{-\mu(|\mathbf{x}-\mathbf{y}|-v_{\text{LR}}|t|)}, \quad (17.44)$$

where $v_{\text{LR}} \sim Ja/\hbar$ is the *Lieb–Robinson velocity* (J is the interaction strength) and $\mu > 0$. Outside the cone $|\mathbf{x} - \mathbf{y}| > v_{\text{LR}}|t|$: correlations are exponentially suppressed. Information cannot propagate faster than v_{LR} .

This is a *derived* causal structure, not a postulated one. It is the inevitable consequence of nearest-neighbour locality: if the Hamiltonian couples only adjacent sites, an excitation at site $\mathbf{x}$ can reach site $\mathbf{y}$ no sooner than $|t| \geq |\mathbf{x} - \mathbf{y}|/v_{\text{LR}}$. In the continuum limit ($a \rightarrow 0$ with v_{LR} held fixed), the maximum propagation speed becomes c , and the Lieb–Robinson cone becomes the light cone of special relativity.

Nearest-neighbour locality therefore gives the first component of Lorentz invariance: a maximum speed and a causal structure. But a light cone is not the full Lorentz group—it gives the causal ordering but not the boost symmetry. The boost symmetry requires the second ingredient.

Ingredient 2: stable Born rendering kills Lorentz violation

A generic lattice Hamiltonian with spacing a produces a dispersion relation for its low-energy excitations:

$$\omega^2(\mathbf{k}) = c^2|\mathbf{k}|^2 + m^2c^4/\hbar^2 + \alpha_4 a^2c^2|\mathbf{k}|^4 + \alpha_6 a^4c^2|\mathbf{k}|^6 + \dots, \quad (17.45)$$

where $c = v_{\text{LR}}$. The first two terms are Lorentz-invariant: they give the relativistic dispersion $E^2 = p^2c^2 + m^2c^4$. The remaining terms ($a^2|\mathbf{k}|^4$, $a^4|\mathbf{k}|^6$, ...) break Lorentz invariance: they depend on the grid spacing a and distinguish the rest frame of the lattice from boosted frames.

Now apply the rendering stability condition from the preface (equation 1.11): Born-layer observables must be independent of a up to $o(a/\lambda)$ corrections. This is the requirement that the physics on the screen does not depend on the mesh of the engine.

The Lorentz-invariant terms ($|\mathbf{k}|^2$, m^2 , the gauge coupling g) can be made a -independent by adjusting finitely many bare parameters $\theta(a)$ —this is renormalization. These terms *survive* the stability condition. They are the relevant and marginal operators of the renormalization group.

The Lorentz-violating terms ($a^2|\mathbf{k}|^4$, $a^4|\mathbf{k}|^6$, ...) depend explicitly on a and cannot be absorbed into finitely many parameters. For the Born output to be a -independent, these terms must *vanish* in the stable limit. They are the irrelevant operators: suppressed by powers of $a|\mathbf{k}| \sim E/E_{\text{Planck}}$, they flow to zero under the renormalization group.

The stability condition does not merely suppress the Lorentz-violating terms; it *selects* the Lorentz-invariant ones. Among all possible dispersion relations on the lattice, only $\omega^2 = c^2|\mathbf{k}|^2 + m^2c^4/\hbar^2$ is stable under changes of a . Lorentz invariance is selected by the stability of the rendering.

The two ingredients combined

	Ingredient	What it gives	Status
1	Nearest-neighbour locality	Maximum speed $c = v_{\text{LR}}$ and a light cone (Lieb–Robinson bound)	Theorem (Lieb–Robinson 1972)
2	Stable Born rendering	Lorentz-violating corrections ($a^2 \mathbf{k} ^4, \dots$) vanish; only Lorentz-invariant terms survive	Renormalization group universality; demonstrated in condensed-matter examples below
1+2	Combined	Full $SO^+(3,1)$ Lorentz invariance at long wavelengths, corrections $\sim (E/E_P)^2$	The mechanism is universality at the stable fixed point

Ingredient 1 gives the causal structure (the light cone) but not the boost symmetry. Ingredient 2 gives the boost symmetry by killing the Lorentz-violating terms. Together: full Lorentz invariance, not as a postulate but as a consequence of the lattice’s locality and the rendering’s stability.

If the lattice spacing is Planckian ($a \sim \ell_P \sim 10^{-35}$ m), the Lorentz-violating corrections are of order $(E/E_P)^2 \sim (E/10^{19} \text{ GeV})^2$. At the LHC ($E \sim 10^4$ GeV): corrections $\sim 10^{-30}$, far below any experimental sensitivity. Current bounds on Lorentz violation (from gamma-ray burst timing, ultra-high-energy cosmic rays, and atomic clock comparisons) are consistent with a Planckian lattice.

Condensed-matter demonstrations

The mechanism is not hypothetical; it is demonstrated in condensed-matter systems where the lattice is physical and the emergent Lorentz invariance is exact at the critical point:

The 2D Ising model at criticality. The lattice has $\mathbb{Z}_2$ symmetry on a square grid (discrete rotations, not continuous). At $T = T_c$: the continuum limit is a free Majorana fermion with the full conformal group—including Lorentz invariance $SO(2,1)$. The lattice anisotropy (the difference between the x - and y -directions) is an irrelevant operator that vanishes at the critical fixed point.

Graphene. The honeycomb lattice has C_{3v} symmetry (threefold rotation). The low-energy excitations are Dirac fermions with linear dispersion $E = v_F|\mathbf{k}|$ at the K and K' points. The effective theory has emergent $SO(2,1)$ Lorentz invariance with v_F playing the role of c . The Lorentz-violating correction (trigonal warping: $k^3 \cos 3\theta$) is irrelevant, suppressed by $(a|\mathbf{k}|)^2$.

The Kitaev honeycomb model. In the gapless phase: Majorana fermions with a Dirac dispersion at the K points and emergent Lorentz invariance. The lattice has only discrete rotational symmetry; the continuous boost symmetry is emergent, selected by the universality of the critical fixed point.

In each case: the lattice has only discrete symmetry; the continuum limit has continuous Lorentz symmetry; and the Lorentz-violating corrections are irrelevant operators killed by the stability of the long-wavelength rendering.

The Fock-Born reading

In the Fock-Born framework, Lorentz invariance is not a postulate about the world. It is a property of the Born-layer rendering, selected by the same stability condition that defines renormalization:

- The **Fock layer** operates on a discrete lattice with spacing a , nearest-neighbour Hamiltonian $H(a)$, and bare parameters $\theta(a)$. The lattice has only discrete symmetries (hypercubic rotations, lattice translations). It is not Lorentz-invariant.
- The **Born layer** renders observables at wavelengths $\lambda \gg a$. The stability condition $\mathcal{O}_{\text{Born}}(\lambda) = \mathcal{O}_{\text{phys}} + o(a/\lambda)$ kills all a -dependent terms in the dispersion relation. The surviving terms are Lorentz-invariant.
- The **speed of light** c is the Lieb–Robinson velocity of the lattice: the maximum propagation speed set by the ratio of interaction strength to lattice spacing. It is not a fundamental constant; it is an architectural property of the engine.
- The **Minkowski metric** $\eta_{\mu\nu}$ is the continuum limit of the lattice’s causal structure, characterised by the light cone $|\Delta\mathbf{x}|^2 - c^2\Delta t^2 = 0$. It is a Born-layer quantity: the metric is what rods and clocks report at $\lambda \gg a$. It is not an input to the Fock layer; it is an output of the stable rendering.

Lorentz invariance is the symmetry of the stable rendering, not the symmetry of the engine. The engine is a lattice. The rendering is Lorentz-invariant. The two are related by the same mechanism—universality at a renormalization-group fixed point—that gives emergent Lorentz invariance to phonons, to graphene’s Dirac fermions, and to the Ising model’s Majorana fermion at criticality.

The Fock-Born framework therefore does not need to postulate $\eta_{\mu\nu}$. It derives $\eta_{\mu\nu}$ from the lattice’s nearest-neighbour locality and the rendering’s stability. The speed of light is the engine’s maximum update speed. Causality is a consequence of the engine’s locality. And Lorentz invariance is a property of the screen, not of the engine that runs behind it.

17.11 Space and time: separate in the engine, packaged in the rendering

The previous subsection showed that Lorentz invariance is a long-distance consequence of nearest-neighbour locality and rendering stability. We now examine what this implies about the relationship between space and time: they are separate objects in the engine, and the Minkowski metric $\eta_{\mu\nu}$ is the formal packaging of this separation into a single rendered object.

The engine separates space and time

The lattice engine has two distinct structures:

Space: a d -dimensional torus $\mathbb{T}^d$ with spacing a . Each lattice site $\mathbf{n} = (n_1, \dots, n_d)$ is a spatial location. The Hamiltonian $H(a)$ couples nearest-neighbour sites.

Time: the parameter t of the Schrödinger equation $h_{t+dt} = e^{-iH(a)dt}h_t$. Time is the engine’s tick: a monotonic parameter that advances the hidden state. It is not a lattice site; it is the parameter of unitary evolution.

Space and time play fundamentally different roles:

Space is the arena (the lattice topology).

Time is the parameter (the evolution generator H).

Fields $\hat{\phi}(\mathbf{n}, t)$ live on spatial sites and evolve in time. One can ask “what is the field at site $\mathbf{n}$ at time t ?” But $\mathbf{n}$ and t are not the same kind of object: $\mathbf{n}$ is a label on the Hilbert space; t is the parameter of the unitary group e^{-iHt} .

This separation is not a limitation of the lattice. It is the structure of quantum mechanics: the Hilbert space $\mathcal{F}_a$ is defined at a given time, and the Hamiltonian generates evolution from one time to the next. The Schrödinger picture, the Heisenberg picture, and the interaction picture

all presuppose this separation.

$\eta_{\mu\nu}$: the ad hoc packaging

The Minkowski metric packages the separate time and space intervals into a single quadratic form:

$$ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2 = \eta_{\mu\nu} dx^\mu dx^\nu. \quad (17.46)$$

This packaging is a rendering-level convenience: it makes the rendering-stable long-distance physics manifestly Lorentz-invariant. But it obscures the engine-level separation.

Where does $\eta_{\mu\nu}$ come from?

The (-1) in $\eta_{00} = -1$ comes from the fact that time evolution is *unitary* (e^{-iHt} , with the i), while spatial propagation is *diffusive/oscillatory* ($e^{ik \cdot x}$, without the i in the exponential of the action). The relative sign between dt^2 and dx^2 reflects the phase-vs-amplitude distinction: time advances the phase (the Fock-layer rotation); space organises the modes (the lattice topology).

The c^2 in front of dt^2 is the ratio $c = a/\Delta t$ (or more precisely, the Lieb–Robinson velocity): the maximum speed at which information propagates on the lattice, set by the nearest-neighbour coupling strength.

The “3+1 signature” $(-, +, +, +)$ is not a deep truth about the unity of space and time. It is the rendering of two separate facts: (1) there is one time direction (one evolution parameter); (2) there are three spatial directions (three lattice axes). Packaging them into $\eta_{\mu\nu} = \text{diag}(-1, +1, +1, +1)$ is efficient notation, not ontology.

$(E, \mathbf{p})$: the same packaging for generators

Energy and momentum are eigenvalues of separate operators:

Energy E : the eigenvalue of the Hamiltonian H , which generates time translations ($t \rightarrow t + dt$).

Momentum $\mathbf{p}$: the eigenvalue of the momentum operator $\mathbf{P}$, which generates spatial translations ($\mathbf{n} \rightarrow \mathbf{n} + \hat{\mu}$).

On the lattice: H and $\mathbf{P}$ are separate operators on $\mathcal{F}_a$. H generates the time tick; $P_\mu = -i\partial/\partial x^\mu$ generates spatial shifts. They commute: $[P_\mu, H] = 0$ (translation invariance in space does not interfere with evolution in time).

The relativistic packaging: combine E and $\mathbf{p}$ into the 4-momentum:

$$p^\mu = (E/c, \mathbf{p}), \quad p_\mu p^\mu = \eta_{\mu\nu} p^\mu p^\nu = -E^2/c^2 + |\mathbf{p}|^2 = -m^2 c^2. \quad (17.47)$$

The mass-shell condition $p^2 = -m^2 c^2$ packages the *separate* eigenvalue conditions $E = \omega(\mathbf{p})$ and $|\mathbf{p}| = |\mathbf{k}|$ into a single Lorentz-invariant equation.

This packaging is a rendering property: it is a consequence of Lorentz invariance (rendering stability at long distances), not of a fundamental unity between energy and momentum. On the lattice: E and $\mathbf{p}$ are separate eigenvalues, computed from separate operators, with separate physical meanings (rate of phase rotation vs. spatial periodicity).

The packaging has practical power—it makes Lorentz covariance manifest and simplifies calculations enormously. But it should not be mistaken for the claim that energy IS momentum in the time direction. Energy is the time-tick generator; momentum is the spatial-shift generator. They happen to be packaged together by the long-distance rendering symmetry.

$g_{\mu\nu}$: the ultimate packaging (and its cost)

General relativity takes the packaging further: the metric $g_{\mu\nu}$ makes space and time *dynamical* and packages them into a single geometric object. The lapse function N , the shift vector N^i , and the spatial metric γ_{ij} are components of $g_{\mu\nu}$:

$$ds^2 = -N^2 c^2 dt^2 + \gamma_{ij}(dx^i + N^i c dt)(dx^j + N^j c dt). \quad (17.48)$$

But even in this maximally packaged form, the physics separates space and time. The ADM formalism (Arnowitt, Deser, Misner, 1962) decomposes the Einstein equations into:

$\gamma_{ij}(t)$: the spatial metric, evolving in time.

$\pi^{ij}(t)$: the conjugate momentum, evolving in time.

N, N^i : Lagrange multipliers (not dynamical, determined by constraints).

The ADM Hamiltonian:

$$H_{\text{ADM}} = \int d^3x (N\mathcal{H} + N^i \mathcal{H}_i), \quad (17.49)$$

generates evolution of γ_{ij} in the time parameter t . The constraints $\mathcal{H} = 0, \mathcal{H}_i = 0$ are the Einstein equations in Hamiltonian form.

The point: even in general relativity, the physics is evolution of spatial data in time. The $\text{Diff}(M)$ packaging (treating space and time as a single manifold) is a rendering symmetry. The engine (the ADM Hamiltonian) still separates them.

Evolution in time is the right physics

At every level of description, the physics is “spatial data evolving in time”:

Level	Spatial data	Time evolution
Condensed matter	$\phi_{\mathbf{n}}$ on lattice $\mathbb{Z}^d$	e^{-iHt} (Schrödinger)
QFT (canonical)	$\hat{\phi}(\mathbf{x})$ on $\mathbb{R}^3$	e^{-iHt} (Schrödinger / Heisenberg)
QFT with Lorentz	Same, but rendering-stable	Same H , Lorentz-covariant
General relativity	$\gamma_{ij}(\mathbf{x})$ on Σ	ADM evolution $\dot{\gamma}_{ij} = \{\gamma_{ij}, H_{\text{ADM}}\}$
Lattice-first	ϕ_n on $\mathbb{T}_a^3$	$e^{-iH(a)dt}$ (engine tick)

Lorentz invariance does not change this: it is a constraint on H (the Hamiltonian must be the zero-component of a 4-vector P^μ with $P^2 = -M^2$), not a replacement of the Hamiltonian picture.

$\text{Diff}(M)$ does not change this: even in GR, the ADM decomposition separates space and time, and the Hamiltonian generates evolution. The “problem of time” (Section ??) is not a failure of the Hamiltonian picture but an artifact of the Diff packaging: promoting the definite time t to a coordinate on the dynamical manifold makes it gauge-dependent.

The packaging hierarchy.

Packaging	What it unifies	Symmetry	Status
None (lattice)	Nothing: t and $\mathbf{n}$ separate	$H(4)$ (lattice)	The engine
$\eta_{\mu\nu}$	dt and dx into ds^2	$SO^+(3,1)$ (Lorentz)	Long-distance rendering
$p^\mu = (E, \mathbf{p})$	E and $\mathbf{p}$ into $p^2 = -m^2$	$SO^+(3,1)$ (same)	Consequence of $\eta_{\mu\nu}$
$g_{\mu\nu}$	dt , dx , and gravity into one metric	$\text{Diff}(M)$	Polished rendering

Each row packages more. Each row adds a symmetry. Each row moves further from the engine. The engine (top row) is the only level where space and time are honestly separate. Every subsequent level is a rendering approximation that trades separation for symmetry.

In one paragraph. Space and time are separate in the engine: space is the lattice, time is the Hamiltonian's parameter. The Minkowski metric $\eta_{\mu\nu}$ packages them into one object, gaining Lorentz invariance (a long-distance smoothness) at the cost of obscuring the separation. The 4-momentum $p^\mu = (E, \mathbf{p})$ is the same packaging for generators. The metric $g_{\mu\nu}$ packages them further, gaining Diff invariance at the cost of the “problem of time.” At every level, the physics is spatial data evolving in time. The packaging (Lorentz, Diff) is a rendering convenience, not a statement about the nature of time.

18 The Scalar Field: Hamiltonian, Correlators, and Schwinger’s Source

The scalar field is the simplest relativistic quantum field theory. We develop it in the Hamiltonian framework—Fock space, creation and annihilation operators, Schrödinger equation—before introducing the path integral in the next section.

Throughout this section, we take the **lattice-first** view: the Fock-layer engine is defined on a spatial lattice with spacing a , and the Hamiltonian $H(a)$ has parameters $\theta(a) = (m^2(a), \lambda(a), Z(a))$ that depend on a . We write differential notation $(\partial_\mu \phi, \int d^3x)$ as shorthand for the lattice expressions $((\phi_{n+\hat{\mu}} - \phi_n)/a, a^3 \sum_n)$ for readability. The lattice provides a natural ultraviolet cutoff: momenta are restricted to the Brillouin zone $|k_\mu| < \pi/a$, and all sums and integrals are finite.

Traditional textbooks call $m^2(a)$ the “bare mass” and m_{phys} (the pole of the propagator) the “physical mass.” We resist this language. The word “bare” suggests a quantity that is stripped or incomplete; in fact $m^2(a)$ is the complete, well-defined mass parameter of the engine at lattice spacing a —no more “bare” than the learning rate of a neural network is “bare.” In the Fock-Born framework:

	Traditional language	Fock-Born language
$m^2(a)$	“Bare mass”	Engine parameter at lattice spacing a
$\lambda(a)$	“Bare coupling”	Engine parameter at lattice spacing a
m_{phys}	“Physical mass”	Born-layer observable (pole of propagator)
σ	“Physical cross section”	Born-layer observable
Renormalisation	“Removing infinities”	Adjusting $\theta(a)$ so that Born-layer observables are a -independent

The engine parameters $\theta(a)$ depend on the lattice spacing; the Born-layer observables do not (to accuracy $o(a/\lambda)$, where λ is the observation scale). This is rendering stability: the physics seen by the observer is insensitive to the engine’s resolution, provided $a \ll \lambda$. Renormalisation is the procedure of adjusting $\theta(a)$ to maintain rendering stability.

18.1 The free scalar field on a lattice

The simplest relativistic quantum field theory is the real scalar field. On a spatial lattice with spacing a and N_s sites per direction, the Fock-layer engine consists of:

The Fock space. One harmonic oscillator per momentum mode $\mathbf{k}$ with $|\mathbf{k}| < \pi/a$:

$$\mathcal{F}_a = \bigoplus_{n=0}^{\infty} \mathcal{H}_{\text{sym}}^{(n)}, \quad [a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'}. \quad (18.1)$$

The free Hamiltonian.

$$H_0(a) = \sum_{|\mathbf{k}| < \pi/a} \omega_{\mathbf{k}} a_{\mathbf{k}}^\dagger a_{\mathbf{k}}, \quad \omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2(a)}, \quad (18.2)$$

after normal ordering. Each mode is an independent oscillator rotating at frequency $\omega_{\mathbf{k}}$. The field operator:

$$\hat{\phi}(\mathbf{x}) = \sum_{|\mathbf{k}| < \pi/a} \frac{1}{\sqrt{2\omega_{\mathbf{k}} V}} [a_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{x}} + a_{\mathbf{k}}^\dagger e^{-i\mathbf{k} \cdot \mathbf{x}}]. \quad (18.3)$$

This is a Hermitian operator on $\mathcal{F}_a$ —a superposition of creation and annihilation operators, one per mode. The sums run over the lattice Brillouin zone $|\mathbf{k}| < \pi/a$: the lattice spacing a provides a natural cutoff.

18.2 Adding interactions

The interacting Hamiltonian:

$$H(a) = H_0(a) + \frac{\lambda(a)}{4!} \int d^3x : \hat{\phi}(\mathbf{x})^4 :, \quad (18.4)$$

where $:\cdots:$ denotes normal ordering and $\lambda(a)$ is the bare coupling at lattice spacing a . Expanding $\hat{\phi}(\mathbf{x})^4$ in creation and annihilation operators generates vertices that transfer amplitude between Fock sectors: $a^\dagger a^\dagger a^\dagger a^\dagger$ (create 4), $a^\dagger a^\dagger a^\dagger a$ (create 3, destroy 1), $\dots$, $aaaa$ (destroy 4). The free motor is diagonal in the Fock basis; the interaction motor mixes sectors.

The interacting vacuum $|\Omega\rangle$ is the ground state of $H(a)$: a single quiet vector whose expansion in the free Fock basis has components in multi-particle sectors—basis-decomposition coefficients, not events.

18.3 Correlators: the central objects of the Fock layer

The *time-ordered correlator* (or Green’s function) is the vacuum expectation value of a time-ordered product of Heisenberg-picture field operators:

$$G^{(n)}(x_1, \dots, x_n) = \langle \Omega | T \{ \hat{\phi}(x_1) \cdots \hat{\phi}(x_n) \} | \Omega \rangle, \quad (18.5)$$

where $\hat{\phi}(x) = e^{iHt} \hat{\phi}(\mathbf{x}) e^{-iHt}$ is the field operator evolved to time t by the full Hamiltonian, and T denotes time ordering (later times to the left).

The two-point function (the propagator):

$$G^{(2)}(x, y) = \langle \Omega | T \{ \hat{\phi}(x) \hat{\phi}(y) \} | \Omega \rangle. \quad (18.6)$$

This is a Fock-layer matrix element: the amplitude for the engine to propagate a field disturbance from y to x (or from x to y , depending on the time ordering). It is *not* a Born-layer observable—no observer submits a “two-point function” to the Born interface. It is an internal property of the engine.

Computing the free propagator from the mode expansion. This is the single most important computation in perturbative quantum field theory: every Feynman diagram is built from free propagators connected at vertices. We derive it step by step from the mode expansion (18.3).

For the free field ($\lambda = 0$), the vacuum is the free vacuum $|0\rangle$ satisfying $a_{\mathbf{k}}|0\rangle = 0$ for all $\mathbf{k}$. Consider $x^0 > y^0$ first (the time-ordered product puts $\hat{\phi}(x)$ to the left):

$$G_0^{(2)}(x, y)|_{x^0 > y^0} = \langle 0 | \hat{\phi}(x) \hat{\phi}(y) | 0 \rangle. \quad (18.7)$$

Substitute the mode expansion:

$$\hat{\phi}(x) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{\sqrt{2\omega_{\mathbf{k}}}} [a_{\mathbf{k}} e^{-ik \cdot x} + a_{\mathbf{k}}^\dagger e^{ik \cdot x}], \quad (18.8)$$

where $k \cdot x = \omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x}$ and $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}$. (We use the continuum limit of (18.3) with $V^{-1} \sum_{\mathbf{k}} \rightarrow \int d^3\mathbf{k}/(2\pi)^3$.)

The product $\hat{\phi}(x)\hat{\phi}(y)$ contains four terms: aa , $aa^\dagger$, $a^\dagger a$, $a^\dagger a^\dagger$. Only the $aa^\dagger$ term survives when sandwiched between $\langle 0|$ and $|0\rangle$, because $a_{\mathbf{k}}|0\rangle = 0$ kills the terms with a on the right, and $\langle 0|a_{\mathbf{k}}^\dagger = 0$ kills terms with $a^\dagger$ on the left:

$$\begin{aligned}\langle 0|\hat{\phi}(x)\hat{\phi}(y)|0\rangle &= \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{d^3\mathbf{k}'}{(2\pi)^3} \frac{1}{2\sqrt{\omega_{\mathbf{k}}\omega_{\mathbf{k}'}}} \underbrace{\langle 0|a_{\mathbf{k}}a_{\mathbf{k}'}^\dagger|0\rangle}_{=(2\pi)^3\delta^3(\mathbf{k}-\mathbf{k}')} e^{-ik\cdot x} e^{ik'\cdot y} \\ &= \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{2\omega_{\mathbf{k}}} e^{-ik\cdot(x-y)}.\end{aligned}\quad (18.9)$$

We used the commutation relation $[a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = (2\pi)^3\delta^3(\mathbf{k}-\mathbf{k}')$ and the fact that $\langle 0|a_{\mathbf{k}}a_{\mathbf{k}'}^\dagger|0\rangle = \langle 0|[a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger]|0\rangle = (2\pi)^3\delta^3(\mathbf{k}-\mathbf{k}')$ (the $a^\dagger a$ piece vanishes on the vacuum).

Similarly, for $x^0 < y^0$: $\langle 0|\hat{\phi}(y)\hat{\phi}(x)|0\rangle = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{2\omega_{\mathbf{k}}} e^{-ik\cdot(y-x)}$.

Combining into the time-ordered propagator. The time-ordered product is:

$$G_0^{(2)}(x, y) = \theta(x^0 - y^0) \langle 0|\hat{\phi}(x)\hat{\phi}(y)|0\rangle + \theta(y^0 - x^0) \langle 0|\hat{\phi}(y)\hat{\phi}(x)|0\rangle. \quad (18.10)$$

Both cases can be combined into a single 4-momentum integral using the identity:

$$\theta(\Delta t) e^{-i\omega\Delta t} + \theta(-\Delta t) e^{i\omega\Delta t} = \int \frac{dk^0}{2\pi} \frac{i e^{-ik^0\Delta t}}{(k^0)^2 - \omega^2 + i\varepsilon}, \quad (18.11)$$

where $\Delta t = x^0 - y^0$ and the $i\varepsilon$ prescription places the poles at $k^0 = \pm(\omega - i\varepsilon)$ —slightly below and above the real axis, ensuring the correct time ordering when the k^0 contour is closed.

To derive (18.11): the integrand has poles at $k^0 = \omega - i\varepsilon$ and $k^0 = -\omega + i\varepsilon$. For $\Delta t > 0$: close the k^0 contour in the lower half-plane (where $e^{-ik^0\Delta t}$ decays), picking up the pole at $k^0 = \omega - i\varepsilon$. Residue: $e^{-i\omega\Delta t}/(2\omega) \cdot (-2\pi i)$, giving $e^{-i\omega\Delta t}/(2\omega)$. For $\Delta t < 0$: close in the upper half-plane, picking up the pole at $k^0 = -\omega + i\varepsilon$. Residue: $e^{i\omega|\Delta t|}/(2\omega)$.

Substituting into the propagator:

$$G_0^{(2)}(x, y) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{2\omega_{\mathbf{k}}} \int \frac{dk^0}{2\pi} \frac{i e^{-ik^0(x^0-y^0)}}{(k^0)^2 - \omega_{\mathbf{k}}^2 + i\varepsilon} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{y})}. \quad (18.12)$$

Using $\omega_{\mathbf{k}}^2 = |\mathbf{k}|^2 + m^2$ and combining the k^0 and $\mathbf{k}$ integrals into a single 4-momentum integral $d^4k = dk^0 d^3\mathbf{k}$:

$$\boxed{G_0^{(2)}(x, y) = \int \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2 + i\varepsilon} e^{-ik\cdot(x-y)},} \quad (18.13)$$

where $k^2 = (k^0)^2 - |\mathbf{k}|^2$ is the Lorentz-invariant 4-momentum squared. In momentum space:

$$\boxed{\tilde{G}_0^{(2)}(k) = \frac{i}{k^2 - m^2 + i\varepsilon}.} \quad (18.14)$$

This is the **Feynman propagator**: the Fourier transform of the free two-point function. Every internal line in a Feynman diagram carries this factor.

What the propagator encodes. The pole at $k^2 = m^2$ (i.e., $k^0 = \pm\omega_{\mathbf{k}}$) is the one-particle state: it corresponds to a single free particle of mass m propagating on shell. The $i\varepsilon$ prescription implements the Feynman boundary condition: positive-frequency modes propagate forward in time, negative-frequency modes propagate backward.

On the lattice. On the lattice with spacing a , the momentum integral is restricted to the Brillouin zone $|k_\mu| < \pi/a$, and the continuum k^2 is replaced by the lattice dispersion:

$$\hat{k}^2 = \frac{4}{a^2} \sum_\mu \sin^2 \frac{k_\mu a}{2}, \quad (18.15)$$

so $\tilde{G}_0^{(2)}(k) = i/(\hat{k}^2 - m^2(a) + i\varepsilon)$. For $|k_\mu|a \ll 1$: $\hat{k}^2 \approx k^2$ and the continuum propagator is recovered. The difference is a lattice artefact of order $(ka)^2$, suppressed in the rendering-stability condition $o(a/\lambda)$.

The four-point function:

$$G^{(4)}(x_1, x_2, x_3, x_4) = \langle \Omega | T \{ \hat{\phi}(x_1) \hat{\phi}(x_2) \hat{\phi}(x_3) \hat{\phi}(x_4) \} | \Omega \rangle. \quad (18.16)$$

This encodes scattering: it contains the information about how two field excitations interact. The LSZ reduction formula (Section II) extracts the scattering amplitude—a Born-layer observable—from this Fock-layer matrix element.

18.4 Connected correlators: genuine multi-point correlations

The full correlator $G^{(n)}$ contains *disconnected* pieces: products of lower correlators that describe independent, uncorrelated sub-processes. For example:

$$G^{(4)}(x_1, x_2, x_3, x_4) = G_{\text{conn}}^{(4)}(x_1, x_2, x_3, x_4) + G^{(2)}(x_1, x_2) G^{(2)}(x_3, x_4) + 2 \text{ permutations.} \quad (18.17)$$

The disconnected pieces describe particles 1–2 propagating independently of particles 3–4. They are products of independent processes that happen to be evaluated in the same vacuum.

The *connected correlator* $G_{\text{conn}}^{(n)}$ subtracts all such factorisations. It measures the genuine n -point correlation that cannot be decomposed into independent sub-processes. The *cluster decomposition principle*: when any subset of points is separated to spatial infinity, $G_{\text{conn}}^{(n)} \rightarrow 0$. Connected correlators vanish at large separation—this is the statement that widely separated experiments are independent.

The connected correlator is the physically meaningful object. It is what appears in the LSZ reduction formula; it is what determines scattering amplitudes; it is what the generating functional $W[J]$ produces. In the language of probability theory: $G^{(n)}$ are moments; $G_{\text{conn}}^{(n)}$ are cumulants.

18.5 Schwinger’s source theory: $Z[J]$, $W[J]$, and $\Gamma[\bar{\phi}]$ from operators

The generating-functional hierarchy $Z[J] \rightarrow W[J] \rightarrow \Gamma[\bar{\phi}]$ is standardly presented in path-integral language. But its origin is operator-based: Julian Schwinger developed the entire framework from 1951 onward without using a path integral. Understanding the operator-side construction clarifies what these objects *are*—they are properties of the Hamiltonian and the vacuum, not of the path-integral measure.

Historical context. In the late 1940s, Schwinger, Tomonaga, and Feynman independently solved the renormalization of QED. Schwinger and Tomonaga worked entirely within the operator formalism; Feynman introduced the path integral and diagrammatic methods. Dyson proved the equivalence of the two approaches.

After this triumph, Schwinger became increasingly dissatisfied with Feynman diagrams. He remarked that they had brought “computation to the masses” but insisted they were “pedagogy, not physics.” His objection was not computational but conceptual: the diagrams narrate Fock-layer bookkeeping in Born-layer language (particles emitted, exchanged, absorbed), inviting the

conflations catalogued in the preface. Schwinger’s alternative was *source theory*: a complete reformulation of QFT in terms of external sources $J(x)$ coupled to the field operator, generating all physical predictions by functional differentiation—without ever drawing a Feynman diagram.

The key paper is Schwinger’s “On the Green’s Functions of Quantized Fields” (1951), in which he introduced the generating functional $Z[J]$ as a vacuum expectation value of a time-ordered exponential of the source-coupled field. The subsequent development— $W[J]$, the effective action $\Gamma[\phi]$ (Jona-Lasinio 1964, DeWitt 1965), the 2PI effective action (Cornwall–Jackiw–Tomboulis 1974)—all follow from Schwinger’s operator framework. The path integral provides an efficient method for *computing* these objects, but the conceptual framework is Schwinger’s, and it is entirely on the Hamiltonian side.

The generating functional $Z[J]$ from operators. Introduce an external classical source $J(x)$ —a prescribed function of spacetime—coupled linearly to the field operator. The modified Hamiltonian:

$$H_J(t) = H - \int d^3\mathbf{x} J(\mathbf{x}, t) \hat{\phi}(\mathbf{x}, t). \quad (18.18)$$

The source J is a handle: it perturbs the engine from outside, allowing us to probe its response.

The generating functional is the vacuum-to-vacuum transition amplitude in the presence of the source:

$$Z[J] = \langle \Omega | T \exp\left(i \int d^4x J(x) \hat{\phi}(x)\right) | \Omega \rangle. \quad (18.19)$$

This is a vacuum matrix element of a time-ordered exponential of field operators—a purely operator-based definition. No path integral. No integration over field configurations. The operator $T \exp\left(i \int J \hat{\phi}\right)$ is defined by its Taylor series:

$$Z[J] = \sum_{n=0}^{\infty} \frac{i^n}{n!} \int d^4x_1 \cdots d^4x_n J(x_1) \cdots J(x_n) \langle \Omega | T \{\hat{\phi}(x_1) \cdots \hat{\phi}(x_n)\} | \Omega \rangle. \quad (18.20)$$

The coefficient of $J(x_1) \cdots J(x_n)$ is exactly the n -point correlator $G^{(n)}(x_1, \dots, x_n)$. Therefore:

$$G^{(n)}(x_1, \dots, x_n) = \frac{1}{i^n} \frac{\delta^n Z[J]}{\delta J(x_1) \cdots \delta J(x_n)} \Big|_{J=0}. \quad (18.21)$$

$Z[J]$ is the generating functional for all correlators: it encodes the complete information of the vacuum and the Hamiltonian in a single object.

$W[J] = \ln Z[J]$: **the connected generating functional.** Taking the logarithm extracts connected correlators:

$$W[J] = \ln Z[J] = \sum_{n=1}^{\infty} \frac{i^n}{n!} \int d^4x_1 \cdots d^4x_n J(x_1) \cdots J(x_n) G_{\text{conn}}^{(n)}(x_1, \dots, x_n). \quad (18.22)$$

The proof is purely algebraic: the logarithm of a moment generating function is the cumulant generating function. This is a theorem of probability theory, not of physics. The connected correlators:

$$G_{\text{conn}}^{(n)}(x_1, \dots, x_n) = \frac{1}{i^n} \frac{\delta^n W[J]}{\delta J(x_1) \cdots \delta J(x_n)} \Big|_{J=0}. \quad (18.23)$$

$W[J]$ is the generating functional for connected correlators. It discards the disconnected pieces automatically. This is why W , not Z , is the object of direct physical interest: the disconnected pieces describe independent processes that have nothing to do with each other.

The classical field $\bar{\phi}$ and the equation of motion. In the presence of the source J , the vacuum expectation value of the field is:

$$\bar{\phi}(x) \equiv \bar{\phi}_J(x) = \frac{\delta W[J]}{\delta J(x)} = \frac{\langle \Omega | T \{ \hat{\phi}(x) e^{i \int J \hat{\phi}} \} | \Omega \rangle}{\langle \Omega | T e^{i \int J \hat{\phi}} | \Omega \rangle}. \quad (18.24)$$

This is the “classical field”: the Born expected value of $\hat{\phi}(x)$ in the presence of the source. When $J = 0$: $\bar{\phi} = \langle \Omega | \hat{\phi}(x) | \Omega \rangle$, the vacuum expectation value. For a scalar field with unbroken symmetry: $\bar{\phi}|_{J=0} = 0$. For a field with spontaneous symmetry breaking: $\bar{\phi}|_{J=0} = v \neq 0$ (the condensate).

The map $J \mapsto \bar{\phi}$ is invertible (for small J): given $\bar{\phi}(x)$, the source $J(x)$ needed to produce it can be found. This invertibility is the precondition for the Legendre transform.

$\Gamma[\bar{\phi}]$: **the effective action.** The Legendre transform:

$$\Gamma[\bar{\phi}] = W[J] - \int d^4x J(x) \bar{\phi}(x), \quad (18.25)$$

where $J(x)$ is understood as a functional of $\bar{\phi}(x)$ through the inversion of (18.24). The effective action $\Gamma[\bar{\phi}]$ is a functional of the classical field $\bar{\phi}$, not of the source J .

The equation of motion for $\bar{\phi}$:

$$\frac{\delta \Gamma[\bar{\phi}]}{\delta \bar{\phi}(x)} = -J(x). \quad (18.26)$$

When $J = 0$: the physical vacuum expectation value $\bar{\phi}$ is a stationary point of Γ —exactly as the classical field is a stationary point of the classical action S . The effective action Γ is the quantum-corrected action: it includes all quantum effects (loops, renormalization, vacuum fluctuations) in a single functional whose stationary point gives the exact vacuum.

The 1PI vertex functions. Functional derivatives of Γ generate the one-particle irreducible (1PI) vertex functions:

$$\Gamma^{(n)}(x_1, \dots, x_n) = \left. \frac{\delta^n \Gamma[\bar{\phi}]}{\delta \bar{\phi}(x_1) \cdots \delta \bar{\phi}(x_n)} \right|_{\bar{\phi}=\bar{\phi}_0}. \quad (18.27)$$

The two-point vertex $\Gamma^{(2)}(x, y)$ is the inverse propagator:

$$\int d^4z \Gamma^{(2)}(x, z) G_{\text{conn}}^{(2)}(z, y) = -i \delta^4(x - y). \quad (18.28)$$

In momentum space: $\Gamma^{(2)}(p) = p^2 - m_0^2 - \Sigma(p)$, where $\Sigma(p)$ is the self-energy. The physical mass is the zero of $\Gamma^{(2)}$:

$$\Gamma^{(2)}(p^2 = m_{\text{phys}}^2) = 0 \iff m_{\text{phys}}^2 = m_0^2 + \text{Re} \Sigma(m_{\text{phys}}^2). \quad (18.29)$$

The four-point vertex $\Gamma^{(4)}$ is the physical four-point coupling at a given momentum configuration.

The key structural property: any connected correlator $G_{\text{conn}}^{(n)}$ can be assembled from 1PI vertices $\Gamma^{(k)}$ connected by full propagators $G_{\text{conn}}^{(2)}$ in a tree structure. The 1PI vertices are the irreducible building blocks; the connected correlators are assembled from them. The effective action Γ encodes the skeleton of the theory.

The hierarchy at a glance:

$$Z[J] \xrightarrow{\ln} W[J] \xrightarrow{\text{Legendre}} \Gamma[\bar{\phi}] \quad (18.30)$$

Functional	Generates	Combinatorics	Analogy
$Z[J]$	All correlators $G^{(n)}$	All diagrams (connected + disconnected)	Moment generating function
$W[J] = \ln Z$	Connected correlators $G_{\text{conn}}^{(n)}$	Connected diagrams only	Cumulant generating function
$\Gamma[\bar{\phi}]$	1PI vertices $\Gamma^{(n)}$	1PI diagrams only	Rate function (large deviations)

Each level strips away reducible structure. Z contains everything including disconnected vacuum bubbles. W removes the disconnected pieces. Γ removes the reducible internal propagators. The physical content is most concentrated in Γ .

The entire hierarchy is on the operator side. Everything above was derived from the operator definition (18.19): a vacuum expectation value of a time-ordered exponential of $\hat{\phi}$. The path integral provides a way to *compute* $Z[J]$ —by writing $Z[J] = \int \mathcal{D}\phi e^{i(S[\phi] + \int J\phi)}$ —but the definition, the structure, and the physical meaning of Z , W , Γ are all Schwinger’s, and they are all on the Hamiltonian side. The path integral is a computational tool that evaluates Schwinger’s generating functional efficiently. Schwinger computed without it; the results are the same.

18.6 The spectral representation: what “a particle” means

Derivation of the Källén–Lehmann representation. Insert a complete set of exact energy-momentum eigenstates $\mathbf{1} = \sum_n |n\rangle \langle n|$ (where n labels all eigenstates of H and $\hat{\mathbf{P}}$) between the two field operators. For $x^0 > y^0$:

$$\langle \Omega | \hat{\phi}(x) \hat{\phi}(y) | \Omega \rangle = \sum_n \langle \Omega | \hat{\phi}(x) | n \rangle \langle n | \hat{\phi}(y) | \Omega \rangle. \quad (18.31)$$

The Heisenberg-picture field $\hat{\phi}(x) = e^{iP \cdot x} \hat{\phi}(0) e^{-iP \cdot x}$ (where $P^\mu = (H, \hat{\mathbf{P}})$) gives:

$$\langle \Omega | \hat{\phi}(x) | n \rangle = \langle \Omega | \hat{\phi}(0) | n \rangle e^{-ip_n \cdot x}, \quad \langle n | \hat{\phi}(y) | \Omega \rangle = \langle n | \hat{\phi}(0) | \Omega \rangle e^{ip_n \cdot y}, \quad (18.32)$$

where p_n^μ is the 4-momentum of state $|n\rangle$ (using $P^\mu | \Omega \rangle = 0$ and $P^\mu | n \rangle = p_n^\mu | n \rangle$). Substituting:

$$\langle \Omega | \hat{\phi}(x) \hat{\phi}(y) | \Omega \rangle = \sum_n | \langle \Omega | \hat{\phi}(0) | n \rangle |^2 e^{-ip_n \cdot (x-y)}. \quad (18.33)$$

Group the states $|n\rangle$ by their invariant mass $\mu^2 = p_n^2 = (p_n^0)^2 - |\mathbf{p}_n|^2$, and integrate over the 3-momentum $\mathbf{p}_n$ (using Lorentz invariance to reduce the sum to an integral over μ^2):

$$\sum_n | \langle \Omega | \hat{\phi}(0) | n \rangle |^2 e^{-ip_n \cdot (x-y)} = \int_0^\infty \frac{d\mu^2}{2\pi} \rho(\mu^2) \int \frac{d^3\mathbf{p}}{(2\pi)^3 2E_{\mathbf{p}}} e^{-ip \cdot (x-y)} \Big|_{p^0=E_{\mathbf{p}}(\mu)}, \quad (18.34)$$

where $E_{\mathbf{p}}(\mu) = \sqrt{|\mathbf{p}|^2 + \mu^2}$ and the spectral function is:

$$\rho(\mu^2) = (2\pi) \sum_n \delta(\mu^2 - m_n^2) | \langle \Omega | \hat{\phi}(0) | n \rangle |^2 \geq 0. \quad (18.35)$$

The inner integral over $\mathbf{p}$ is exactly the free propagator of a particle of mass μ , so after time-ordering:

$$G^{(2)}(x, y) = \int_0^\infty \frac{d\mu^2}{2\pi} \rho(\mu^2) G_{0,\mu}^{(2)}(x, y), \quad (18.36)$$

where $G_{0,\mu}^{(2)}$ is the free Feynman propagator with mass μ . In momentum space (using $\tilde{G}_{0,\mu}^{(2)}(k) = i/(k^2 - \mu^2 + i\varepsilon)$):

$$\tilde{G}^{(2)}(k) = \int_0^\infty d\mu^2 \frac{\rho(\mu^2)}{k^2 - \mu^2 + i\varepsilon}. \quad (18.37)$$

This is the Källén–Lehmann representation. It is exact (non-perturbative) and follows from three ingredients alone: Lorentz invariance, completeness of the energy eigenstates, and positivity of the spectral function ($\rho \geq 0$ because it is a sum of squared matrix elements).

The physical content: the full interacting propagator is a superposition of free propagators with mass μ , weighted by $\rho(\mu^2)$ —the probability that the field operator, applied to the vacuum, produces a state of invariant mass μ .

In momentum space, this gives the Källén–Lehmann representation:

$$G^{(2)}(p) = \int_0^\infty d\mu^2 \frac{\rho(\mu^2)}{p^2 - \mu^2 + i\varepsilon}, \quad (18.38)$$

where the spectral function:

$$\rho(\mu^2) = \sum_n |\langle \Omega | \hat{\phi}(0) | n \rangle|^2 (2\pi) \delta(\mu^2 - m_n^2) \quad (18.39)$$

is a sum of delta functions at the squared masses m_n^2 of all states that $\hat{\phi}$ can create from the vacuum, weighted by the squared matrix element $|\langle \Omega | \hat{\phi} | n \rangle|^2$.

The spectral function has a very specific structure:

The one-particle pole. If there is a stable one-particle state $|\mathbf{p}\rangle$ with mass m_{phys} , the spectral function has an isolated delta: $\rho(\mu^2) = Z \delta(\mu^2 - m_{\text{phys}}^2) + \rho_{\text{cont}}(\mu^2)$, where $Z = |\langle \Omega | \hat{\phi}(0) | \mathbf{p} = 0 \rangle|^2$ is the wave-function renormalisation. The propagator:

$$G^{(2)}(p) = \frac{iZ}{p^2 - m_{\text{phys}}^2 + i\varepsilon} + \int_{4m^2}^\infty d\mu^2 \frac{\rho_{\text{cont}}(\mu^2)}{p^2 - \mu^2 + i\varepsilon}. \quad (18.40)$$

The pole at $p^2 = m_{\text{phys}}^2$ IS the particle. A particle, in this framework, is not a tiny billiard ball. It is an isolated pole of the propagator—a discrete contribution to the spectral function—a state in the spectrum of H that the field operator can excite from the vacuum. The mass m_{phys} is a Born-layer observable (what a spectrometer reports). The residue Z is a Fock-layer normalization factor.

The multi-particle continuum. Above the two-particle threshold $\mu^2 \geq 4m_{\text{phys}}^2$, the spectral function is continuous: it describes multi-particle states (two particles, three particles, ...) that the field operator can also excite. The continuum is the background against which the pole stands out.

The resonance. An unstable particle (e.g., the Z boson, the Higgs boson) does not produce a real pole on the physical sheet. Instead, it produces a pole at complex $p^2 = m^2 - im\Gamma$, shifted into the second Riemann sheet:

$$G^{(2)}(p) \sim \frac{iZ}{p^2 - m^2 + im\Gamma}. \quad (18.41)$$

The real part m is the resonance mass; Γ is the decay width; $\tau = 1/\Gamma$ is the lifetime. The Breit–Wigner distribution $\sigma(E) \propto 1/[(E^2 - m^2)^2 + m^2\Gamma^2]$ is the Born-layer cross section near the resonance—the peak shape that experimentalists fit to extract m and Γ .

Derivation of the inversion relation. The relation $\Gamma^{(2)} = (G_{\text{conn}}^{(2)})^{-1}$ follows from the chain rule applied to the Legendre transform.

From (18.24): $\bar{\phi}(x) = \delta W / \delta J(x)$. From (18.26): $J(x) = -\delta \Gamma / \delta \bar{\phi}(x)$. These are inverse maps: $J \mapsto \bar{\phi}$ and $\bar{\phi} \mapsto J$. Apply the chain rule to the identity $\bar{\phi}(x) = \delta W[J[\bar{\phi}]] / \delta J(x)$:

$$\delta^4(x-y) = \frac{\delta \bar{\phi}(x)}{\delta \bar{\phi}(y)} = \int d^4 z \frac{\delta \bar{\phi}(x)}{\delta J(z)} \frac{\delta J(z)}{\delta \bar{\phi}(y)} = \int d^4 z \frac{\delta^2 W}{\delta J(x) \delta J(z)} \frac{\delta^2 \Gamma}{\delta \bar{\phi}(z) \delta \bar{\phi}(y)} \cdot (-1). \quad (18.42)$$

(The (-1) comes from $J = -\delta \Gamma / \delta \bar{\phi}$.) Recognising $\delta^2 W / \delta J(x) \delta J(z)|_{J=0} = i G_{\text{conn}}^{(2)}(x, z)$ (from (18.23)) and $\delta^2 \Gamma / \delta \bar{\phi}(z) \delta \bar{\phi}(y)|_{\bar{\phi}=\bar{\phi}_0} = \Gamma^{(2)}(z, y)$:

$$\int d^4 z G_{\text{conn}}^{(2)}(x, z) \Gamma^{(2)}(z, y) = -i \delta^4(x-y). \quad (18.43)$$

In momentum space (convolution becomes multiplication):

$$\tilde{G}_{\text{conn}}^{(2)}(p) \tilde{\Gamma}^{(2)}(p) = -i. \quad (18.44)$$

Therefore $\tilde{\Gamma}^{(2)}(p) = -i / \tilde{G}_{\text{conn}}^{(2)}(p)$: the two-point vertex function is the inverse of the connected propagator. For the free field: $\tilde{G}_0^{(2)}(p) = i / (p^2 - m^2)$, so $\tilde{\Gamma}_0^{(2)}(p) = -i \cdot (p^2 - m^2) / i = p^2 - m^2$ —the Klein–Gordon operator in momentum space. For the interacting field: $\tilde{\Gamma}^{(2)}(p) = p^2 - m_0^2 - \Sigma(p)$, where $\Sigma(p)$ is the self-energy (the sum of all 1PI diagrams with two external legs). The physical mass satisfies $\tilde{\Gamma}^{(2)}(m_{\text{phys}}^2) = 0$, i.e., $m_{\text{phys}}^2 = m_0^2 + \text{Re } \Sigma(m_{\text{phys}}^2)$.

18.7 The interaction picture and the Dyson series

Dirac's split. Write $H = H_0 + H_{\text{int}}$. In the interaction picture: operators evolve under H_0 , states evolve under H_{int} . The interaction-picture field: $\hat{\phi}_I(x) = e^{iH_0 t} \hat{\phi}(\mathbf{x}) e^{-iH_0 t}$ is the free field (it satisfies the free Klein–Gordon equation). The time-evolution operator in the interaction picture:

$$U_I(t, t_0) = T \exp\left(-i \int_{t_0}^t dt' H_{\text{int}}^I(t')\right), \quad (18.45)$$

where $H_{\text{int}}^I(t) = e^{iH_0 t} H_{\text{int}} e^{-iH_0 t}$.

The S -matrix. $S = \lim_{T \rightarrow \infty} U_I(T, -T) = T \exp\left(-i \int_{-\infty}^{\infty} dt H_{\text{int}}^I\right)$. The scattering amplitude:

$$\langle f | S | i \rangle = \langle f | T \exp\left(-i \int dt H_{\text{int}}^I\right) | i \rangle, \quad (18.46)$$

where $|i\rangle, |f\rangle$ are free-particle Fock states. The Born probability $P(i \rightarrow f) = |\langle f | S | i \rangle|^2$ is the cross section.

The Gell-Mann–Low formula. The interacting vacuum $|\Omega\rangle$ is related to the free vacuum $|0\rangle$ by:

$$|\Omega\rangle = \lim_{T \rightarrow \infty (1-i\varepsilon)} \frac{e^{-iHT} |0\rangle}{\langle 0 | e^{-iHT} | 0 \rangle}. \quad (18.47)$$

The $i\varepsilon$ prescription projects onto the ground state: at large T , $e^{-iHT} |0\rangle$ is dominated by $|\Omega\rangle$ because all excited states acquire a factor $e^{-\varepsilon E_n T}$ that kills them. This connects the interacting correlators to free-field expectation values:

$$G^{(n)}(x_1, \dots, x_n) = \frac{\langle 0 | T \{ \hat{\phi}_I(x_1) \cdots \hat{\phi}_I(x_n) e^{-i \int H_{\text{int}}^I} \} | 0 \rangle}{\langle 0 | T e^{-i \int H_{\text{int}}^I} | 0 \rangle}. \quad (18.48)$$

The numerator is a free-field vacuum expectation value of interaction-picture operators; the denominator cancels the disconnected vacuum bubbles. This formula is the starting point for

perturbation theory: Taylor-expand $e^{-i \int H_{\text{int}}^I}$ in powers of λ , evaluate by Wick's theorem using the free propagator $G_0^{(2)}(x, y) = \langle 0 | T \{ \hat{\phi}_I(x) \hat{\phi}_I(y) \} | 0 \rangle$, and organise the terms as Feynman diagrams.

The Dyson series produces the same diagrams as the path integral. Each term in the Taylor expansion of (18.48) is a product of free propagators contracted according to Wick's theorem. The diagrammatic rules—propagator lines for $G_0^{(2)}$, vertices for $-i\lambda/4!$, momentum-conservation delta functions—are identical to those derived from the path integral. The operator formalism and the path integral are two computational methods for the same Fock-layer matrix elements. The Feynman diagram is the shared notation; the physics is in $Z[J]$, $W[J]$, $\Gamma[\bar{\phi}]$, all of which are Schwinger's operator-side constructions.

19 How Is a Particle Observed?

The previous section built the Fock-layer engine for the scalar field: the Hamiltonian $H(a)$, the correlators $G^{(n)}$, the connected correlators $G_{\text{conn}}^{(n)}$, and the generating-functional hierarchy $Z[J] \rightarrow W[J] \rightarrow \Gamma[\phi]$. None of these are Born-layer observables. They are internal properties of the engine—amplitudes, matrix elements, vertex functions—that the engine computes but the observer never sees directly.

This section asks the question that textbooks rarely make explicit: how does a particle become an observable? Through what Born-layer interface does the propagator’s pole become a mass measurement, the S -matrix element become a cross section, the spectral function become a resonance peak?

The answer in every case is the same: the Born rule $P(\lambda) = \|P_\lambda h_t\|^2$, applied to a macroscopic detector observable with enormous degeneracy, converts a Fock-layer matrix element into a Born-layer event.

19.1 A detector is a macroscopic Born-layer system

A particle detector—a scintillator, a wire chamber, a calorimeter, a photomultiplier—is a macroscopic system with $\sim 10^{23}$ degrees of freedom. When energy is deposited in the detector, the macroscopic observable (the current, the light flash, the temperature rise) undergoes a transition. This transition is a Born-layer event: the eigenvalue of a highly degenerate macroscopic observable rendered to the observer.

The degeneracy is enormous: the macroscopic “click” is an eigenvalue shared by $\sim e^{S/k_B}$ microstates (all the microscopic configurations that produce the same current or the same light flash). The Born distribution for this eigenvalue is near-delta: $P(\text{click}) \approx 1$ or $P(\text{click}) \approx 0$. The detector either fires or it doesn’t. When it fires, the Born-layer has rendered one eigenvalue from the Fock-layer state.

The particle is not “seen.” What is rendered is an eigenvalue of a detector observable—a current in a wire, a flash in a crystal. The word “particle” is the physicist’s interpretation: a pattern of detector clicks consistent with an isolated pole of the propagator at mass m_{phys} , moving along a trajectory consistent with $E^2 = p^2 + m^2$.

19.2 Scattering as a Born probability

A scattering experiment has three stages:

Preparation (Fock layer). The accelerator prepares an initial Fock state. To an excellent approximation, this is a two-particle state $|i\rangle = |\mathbf{p}_1, \mathbf{p}_2\rangle$ with definite momenta. The preparation is a Fock-layer operation: setting the initial condition $h_0 = |i\rangle$.

Evolution (Fock layer). The Hamiltonian H evolves the state: $h_T = e^{-iHT} h_0$. In the interaction picture, this is $S|i\rangle$ where $S = T \exp(-i \int H_{\text{int}}^I dt)$. The state $S|i\rangle$ is a superposition over all possible final states—all momentum configurations, all particle numbers:

$$S|i\rangle = \sum_f \langle f| S|i\rangle |f\rangle. \quad (19.1)$$

This is the engine running. No observation has occurred.

Detection (Born layer). The detector array submits a final-state observable. The Born probability:

$$\boxed{P(i \rightarrow f) = |\langle f| S|i\rangle|^2.} \quad (19.2)$$

This is the cross section σ : an event rate per unit flux, a Born-layer number. It is what the experiment reports. Every cross section measurement is a Born probability: the squared modulus of a Fock-layer S -matrix element.

The distinction is clean:

Fock-layer quantity	Born-layer observable
S -matrix element $\langle f S i \rangle \in \mathbb{C}$	Cross section $\sigma \propto \langle f S i \rangle ^2 \in \mathbb{R}_{\geq 0}$
Amplitude (complex, interferes)	Probability (real, positive, additive)
Computed by Dyson series or path integral	Measured by counting detector clicks
Contains phase information	Phase information lost

19.3 The LSZ reduction formula: from correlators to cross sections

The Lehmann–Symanzik–Zimmermann (LSZ) formula is the bridge from Fock-layer correlators to Born-layer cross sections. It extracts the S -matrix element from the connected correlator by amputating the external propagators and putting the external momenta on shell:

$$\langle f | S | i \rangle = \prod_{j \in \text{ext}} (-i(p_j^2 - m_{\text{phys}}^2)/\sqrt{Z}) \tilde{G}_{\text{conn}}^{(n)}(p_1, \dots, p_n) \Big|_{p_j^2 = m_{\text{phys}}^2}, \quad (19.3)$$

where $\tilde{G}_{\text{conn}}^{(n)}$ is the Fourier transform of the connected n -point correlator and Z is the wave-function renormalisation from the Källén–Lehmann representation (18.40).

The structure of LSZ:

$$G_{\text{conn}}^{(n)} \xrightarrow{\text{Fourier}} \tilde{G}_{\text{conn}}^{(n)}(p_1, \dots, p_n) \xrightarrow{\text{amputate}} \mathcal{M}(p_1, \dots, p_n) \xrightarrow{|\cdot|^2} \sigma. \quad (19.4)$$

The first two steps are Fock-layer operations (manipulating matrix elements of the Hamiltonian). The last step—taking the squared modulus—is the Born rule. LSZ is the exit gate from the Fock layer to the Born layer.

Why connected? Only the connected correlator contributes to scattering. The disconnected pieces describe particles that pass through each other without interacting—they contribute to the forward-scattering “1” in $S = 1 + iT$ but not to the transition amplitude T . The scattering cross section comes from $|\langle f | iT | i \rangle|^2$, which requires $G_{\text{conn}}^{(n)}$.

Why amputate? The external propagators $G^{(2)}(p_j)$ in $\tilde{G}_{\text{conn}}^{(n)}$ have poles at $p_j^2 = m_{\text{phys}}^2$. These poles are the one-particle states of the asymptotic particles. Amputation removes these poles, leaving the interaction kernel $\mathcal{M}$ —the part due to the dynamics, not the free propagation. The factor $(p_j^2 - m_{\text{phys}}^2)$ cancels the pole; dividing by $\sqrt{Z}$ normalises the external states correctly.

19.4 The anatomy of $2 \rightarrow 2$ scattering

The simplest non-trivial scattering process is $2 \rightarrow 2$: two particles in, two particles out. The kinematics of this process is governed by three Lorentz-invariant variables and three topologically distinct exchange channels. Both are Fock-layer constructs—organisational tools for the perturbative expansion of $\mathcal{M}$ —but they determine the structure of every Born-layer cross section.

The Mandelstam variables

Label the four external momenta as p_1, p_2 (incoming) and p_3, p_4 (outgoing), with all momenta defined as ingoing (so p_3 and p_4 have negative energy for outgoing particles). Momentum conservation: $p_1 + p_2 + p_3 + p_4 = 0$.

The three Mandelstam variables:

$$\boxed{s = (p_1 + p_2)^2, \quad t = (p_1 + p_3)^2, \quad u = (p_1 + p_4)^2,} \quad (19.5)$$

with the constraint $s + t + u = m_1^2 + m_2^2 + m_3^2 + m_4^2$ (from momentum conservation and $p_i^2 = m_i^2$). For equal-mass particles: $s + t + u = 4m^2$.

Physical interpretation. In the centre-of-mass (CM) frame of particles 1 and 2:

$s = (p_1 + p_2)^2 = E_{\text{CM}}^2$: the total CM energy squared. For a collider experiment, $\sqrt{s}$ is the available energy for producing new particles.

$t = (p_1 + p_3)^2$: related to the momentum transfer. In the CM frame with scattering angle θ (the angle between $\mathbf{p}_1$ and $-\mathbf{p}_3$):

$$t = -\frac{s - 4m^2}{2} (1 - \cos \theta) \quad (\text{equal masses, CM frame}). \quad (19.6)$$

Forward scattering ($\theta = 0$): $t = 0$. Backward scattering ($\theta = \pi$): $t = -(s - 4m^2)$. The variable t controls the angular distribution: the differential cross section $d\sigma/dt$ is the most natural Lorentz-invariant measure of the angular dependence.

$u = (p_1 + p_4)^2$: the “crossed” momentum transfer (u is t with particles 3 and 4 exchanged).

The three channels

In a theory with a single interaction vertex (like ϕ^4), the tree-level amplitude for $2 \rightarrow 2$ scattering is a sum of three topologically distinct Feynman diagrams, called *channels*:

The s -channel (“annihilation” or “resonance” channel):

$$\mathcal{M}_s = \frac{(-i\lambda)^2 \cdot i}{s - m^2 + i\varepsilon} \quad (\text{two vertices, one propagator carrying momentum } q^2 = s). \quad (19.7)$$

Particles 1 and 2 “annihilate” into a single intermediate propagator, which then “produces” particles 3 and 4. The intermediate propagator carries the total CM momentum $q = p_1 + p_2$, with invariant mass $q^2 = s$.

If $s = M^2$ for some particle of mass M in the theory: the s -channel propagator has a pole, and the amplitude is resonantly enhanced. This is a *resonance*: the scattering proceeds through the formation and decay of an intermediate particle. The Breit–Wigner peak in the cross section $\sigma(s) \propto 1/((s - M^2)^2 + M^2\Gamma^2)$ is the Born-layer signature of an s -channel resonance.

Examples: $e^+e^- \rightarrow Z \rightarrow \mu^+\mu^-$ (the Z resonance at $\sqrt{s} = 91.2 \text{ GeV}$); $\pi^+\pi^- \rightarrow \rho \rightarrow \pi^+\pi^-$ (the ρ resonance at $\sqrt{s} = 770 \text{ MeV}$); $gg \rightarrow H \rightarrow \gamma\gamma$ (Higgs production and decay at the LHC).

The t -channel (“exchange” channel):

$$\mathcal{M}_t = \frac{(-i\lambda)^2 \cdot i}{t - m^2 + i\varepsilon} \quad (\text{propagator carries momentum } q^2 = t). \quad (19.8)$$

Particle 1 “emits” an intermediate propagator, which is “absorbed” by particle 2. The intermediate propagator carries the momentum transfer $q = p_1 + p_3$, with $q^2 = t < 0$ for physical

scattering (spacelike momentum transfer). Since $t < 0$: the t -channel propagator never goes on shell (no resonance). The t -channel describes *exchange*: the force between two particles mediated by the exchange of the intermediate propagator.

At low momentum transfer ($|t| \ll m^2$): $\mathcal{M}_t \approx -\lambda^2/(m^2 - t) \approx -\lambda^2/m^2$, and the Fourier transform to position space gives the Yukawa potential $V(r) \propto e^{-mr}/r$. At $m = 0$ (massless exchange): the Coulomb potential $V(r) \propto 1/r$.

Examples: electron-electron scattering via photon exchange (t -channel photon); nucleon-nucleon scattering via pion exchange (t -channel pion, the origin of the nuclear force).

The u -channel (“crossed exchange” channel):

$$\mathcal{M}_u = \frac{(-i\lambda)^2 \cdot i}{u - m^2 + i\varepsilon} \quad (\text{propagator carries momentum } q^2 = u). \quad (19.9)$$

The u -channel is the t -channel with the two final-state particles exchanged ($3 \leftrightarrow 4$). For identical particles (like $\phi\phi \rightarrow \phi\phi$ in ϕ^4 theory), the t - and u -channels are related by Bose symmetry: $\mathcal{M}$ must be symmetric under $p_3 \leftrightarrow p_4$ (equivalently $t \leftrightarrow u$).

The full tree-level amplitude. For ϕ^4 theory at tree level: there is only a single contact vertex ($-i\lambda$), with no propagator. The three channels appear at one loop. For a theory with a cubic coupling $g\phi^3$, the tree-level amplitude is the sum:

$$\mathcal{M}_{\text{tree}} = \mathcal{M}_s + \mathcal{M}_t + \mathcal{M}_u = (-ig)^2 \left[\frac{i}{s - m^2} + \frac{i}{t - m^2} + \frac{i}{u - m^2} \right]. \quad (19.10)$$

The three terms have different physical content: $\mathcal{M}_s$ produces resonances (peaks in σ at $s = m^2$); $\mathcal{M}_t$ and $\mathcal{M}_u$ produce exchange forces (angular distributions controlled by t and u). The full amplitude is the coherent sum: the three channels interfere.

The physical content of each channel

Channel	Variable	Propagator q^2	Physical process	Born-layer signature
s	$(p_1+p_2)^2$	$q^2 = s > 0$ (timelike)	Resonance: form and decay of intermediate particle	Breit–Wigner peak in $\sigma(s)$
t	$(p_1+p_3)^2$	$q^2 = t < 0$ (space-like)	Exchange: force mediated by virtual particle	Angular distribution $d\sigma/dt$
u	$(p_1+p_4)^2$	$q^2 = u < 0$ (space-like)	Crossed exchange: t -channel with final states swapped	Backward-peaked $d\sigma/d\cos\theta$

Crossing symmetry

The amplitude $\mathcal{M}(p_1, p_2, p_3, p_4)$, viewed as a function of the complex variables s, t, u (subject to $s + t + u = 4m^2$), describes three different physical processes depending on which variables are in the physical region:

s-channel process ($s > 4m^2, t, u < 0$): $\phi(p_1) + \phi(p_2) \rightarrow \phi(-p_3) + \phi(-p_4)$ (direct scattering).

t-channel process ($t > 4m^2, s, u < 0$): $\phi(p_1) + \bar{\phi}(-p_3) \rightarrow \bar{\phi}(-p_2) + \phi(p_4)$ (scattering with particles 2 and 3 exchanged and replaced by antiparticles).

u-channel process ($u > 4m^2$, $s, t < 0$): $\phi(p_1) + \bar{\phi}(-p_4) \rightarrow \bar{\phi}(-p_2) + \phi(p_3)$ (particles 2 and 4 exchanged).

Crossing symmetry is the statement that a single analytic function $\mathcal{M}(s, t, u)$ describes all three processes: the physical amplitude for the t -channel process is obtained by analytically continuing the s -channel amplitude from the region $s > 4m^2$ to the region $t > 4m^2$.

Crossing symmetry is a deep property of quantum field theory. It follows from:

- (i) *Locality*: the fields commute at spacelike separation.
- (ii) *Analyticity*: the amplitude $\mathcal{M}(s, t)$ is analytic in s and t except for poles (bound states and resonances) and branch cuts (multiparticle thresholds).
- (iii) *CPT invariance*: particle and antiparticle have the same mass.

Together, these imply that a particle in the initial state with momentum p can be “crossed” to an antiparticle in the final state with momentum $-p$, and the amplitude is related by analytic continuation.

Worked example: ϕ^4 at one loop

At one loop in ϕ^4 theory, the vertex correction to the four-point function has three contributions (one per channel):

$$\delta\Gamma^{(4)}(s, t, u) = I(s) + I(t) + I(u), \quad (19.11)$$

where the loop integral in the s -channel:

$$I(s) = \frac{\lambda^2}{2} \int_{|k| < \pi/a} \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2} \frac{i}{(p_1 + p_2 - k)^2 - m^2}. \quad (19.12)$$

This is the “bubble” integral: a loop of two propagators carrying momenta k and $p_1 + p_2 - k$. The external momentum flowing through the bubble is $q = p_1 + p_2$, with $q^2 = s$. The integrals $I(t)$ and $I(u)$ are the same function with s replaced by t or u .

Evaluating (in the Euclidean region, with Wick rotation):

$$I(s) = \frac{\lambda^2}{32\pi^2} \left[\ln \frac{\pi^2/a^2}{\mu^2} + f(s/m^2) + \dots \right], \quad (19.13)$$

where μ is the renormalisation scale and $f(s/m^2)$ is a finite function:

$$f(x) = 2 - \sqrt{1 - 4/x} \ln \frac{\sqrt{1 - 4/x} + 1}{\sqrt{1 - 4/x} - 1} \quad (\text{for } x > 4, \text{ i.e., } s > 4m^2). \quad (19.14)$$

The logarithmic divergence $\ln(\pi^2/(a^2\mu^2))$ is the same in all three channels: it is absorbed into the running coupling $\lambda(\mu)$. The finite function $f(s/m^2)$ is channel-dependent: it produces the s -dependent (or t -, u -dependent) structure of the amplitude after renormalisation.

The imaginary part and unitarity. For $s > 4m^2$ (above the two-particle threshold): $f(s/m^2)$ develops an imaginary part:

$$\text{Im } I(s) = \frac{\lambda^2}{32\pi} \sqrt{1 - \frac{4m^2}{s}} \quad (\text{for } s > 4m^2). \quad (19.15)$$

This is the optical theorem: the imaginary part of the forward scattering amplitude equals the total cross section (summed over all final states). The imaginary part arises because the intermediate state in the s -channel loop can go on shell ($k^2 = m^2$ and $(q - k)^2 = m^2$ simultaneously when $s > 4m^2$): the virtual process becomes a real process.

For $t < 0$ and $u < 0$ (the spacelike exchanges): $I(t)$ and $I(u)$ are real. No imaginary part, no on-shell intermediate state. The exchange channels do not cut.

The Fock-Born reading

The three channels are Fock-layer organisational tools: they classify the terms in the perturbative expansion of $\mathcal{M}$ by the momentum flow through the internal propagator. The “annihilation” (s -channel), “exchange” (t -channel), and “crossed exchange” (u -channel) are not Born-layer events: no observer witnesses the formation of an intermediate particle or the exchange of a virtual boson. The channels are computational decompositions of one matrix element into three topological contributions.

The Born-layer observables are:

Resonances: the Breit–Wigner peak in $\sigma(s)$ near $s = M^2$ is a Born-layer signature of the s -channel pole. The peak position (M) and width (Γ) are Born-layer numbers extracted from event-rate histograms.

Angular distributions: the differential cross section $d\sigma/dt$ (or $d\sigma/d\cos\theta$) encodes the relative contributions of the t - and u -channels. Forward peaking ($d\sigma/dt$ large at $|t| \rightarrow 0$) signals t -channel exchange; backward peaking signals u -channel exchange. The angular distribution is a Born-layer observable: a histogram of scattering angles.

Crossing relations: the Born-layer cross sections for the three physical processes (s -channel, t -channel, u -channel reactions) are related by analytic continuation of the same Fock-layer amplitude $\mathcal{M}(s, t, u)$. This is a non-trivial prediction: the cross section for $e^+e^- \rightarrow \mu^+\mu^-$ (s -channel) is related by crossing to the cross section for $e^-\mu^- \rightarrow e^-\mu^-$ (t -channel) — two different experiments, connected by the analytic structure of a single Fock-layer function.

19.5 What is a particle? What is a resonance?

The Källén–Lehmann spectral representation (18.38) provides a precise, non-metaphorical definition of “particle”:

A stable particle is an isolated pole of $G^{(2)}(p)$ on the real p^2 -axis:

$$G^{(2)}(p) \xrightarrow{p^2 \rightarrow m_{\text{phys}}^2} \frac{iZ}{p^2 - m_{\text{phys}}^2 + i\varepsilon}. \quad (19.16)$$

The mass m_{phys} is the pole position—an eigenvalue of H in the one-particle sector. Equivalently, it is the zero of the inverse propagator: $\Gamma^{(2)}(p^2 = m_{\text{phys}}^2) = 0$.

A resonance (unstable particle) is a pole of $G^{(2)}(p)$ in the complex p^2 -plane, on the second Riemann sheet:

$$p_{\text{pole}}^2 = m^2 - im\Gamma, \quad (19.17)$$

where m is the resonance mass and Γ is the total decay width. The lifetime is $\tau = 1/\Gamma$. Near the resonance, the cross section has the Breit–Wigner shape:

$$\sigma(E) \propto \frac{1}{(E^2 - m^2)^2 + m^2\Gamma^2}. \quad (19.18)$$

The peak position gives m ; the peak width gives Γ . Both are Born-layer observables: numbers extracted from the event-rate distribution $\sigma(E)$, which is itself a Born probability.

The diagnostic:

Quantity	Layer	What it is
Propagator $G^{(2)}(x, y)$	Fock	Matrix element of $T\{\hat{\phi}(x)\hat{\phi}(y)\}$
Spectral function $\rho(\mu^2)$	Fock	Sum of $ \langle \Omega \hat{\phi} n \rangle ^2 \delta(\mu^2 - m_n^2)$
Pole position m_{phys}	Born	What the spectrometer reports
Self-energy $\Sigma(p)$	Fock	$\Gamma^{(2)}(p) - (p^2 - m_0^2)$
Decay width Γ	Born	$1/\tau$, fitted from Breit–Wigner
Scattering amplitude $\mathcal{M}$	Fock	Amputated connected correlator
Cross section σ	Born	$ \mathcal{M} ^2 \times (\text{phase space})$
Breit–Wigner peak	Born	$\sigma(E)$ near $E \approx m$

The propagator is Fock-layer. The pole position is extracted from the propagator by analytic continuation—still a Fock-layer operation. The mass becomes a Born-layer observable when a detector measures it: the spectrometer submits an energy observable, the Born rule renders the eigenvalue, and the pattern of rendered eigenvalues is fitted to $E^2 = p^2 + m_{\text{phys}}^2$.

Narrating the propagator as “a particle travelling from y to x ” is the conflation: it pushes a Born-layer concept (a particle detected at a location) into a Fock-layer matrix element (the propagator, which is an amplitude, not a trajectory).

19.6 What experimentalists actually report

Experimental publications report Born-layer quantities:

Cross sections $\sigma(E)$: event rates as a function of centre-of-mass energy. These are Born probabilities $|\langle f | S | i \rangle|^2$, integrated over detector acceptance and summed over unmeasured final states (a partial trace—the Born rule with degenerate observables from section 9).

Masses and widths: extracted from fits to the Breit–Wigner shape (19.18) in the cross-section data. The Particle Data Group reports m and Γ with uncertainties: these are Born-layer observables derived from Born-layer event rates.

Branching ratios: $\text{BR}(X \rightarrow Y) = \Gamma(X \rightarrow Y)/\Gamma_{\text{total}}$. Each partial width $\Gamma(X \rightarrow Y)$ is a Born-layer quantity (the rate of a specific decay channel); the ratio is Born-layer over Born-layer.

Differential distributions: $d\sigma/d\Omega$, $d\sigma/dE$, rapidity distributions, transverse-momentum spectra. All are Born probabilities differentiated with respect to kinematic variables defined by the detector geometry.

The Fock-layer quantities— $\mathcal{M}$, Σ , $\Gamma^{(n)}$, $G_{\text{conn}}^{(n)}$ —never appear in experimental papers. They appear in theory papers, which compute them and predict the Born-layer outputs. The comparison between theory and experiment is always at the Born layer: predicted σ versus measured σ .

19.7 The catalog of Born-layer observables for the scalar field

Observable	Fock-layer source	Bridge	Born-layer output
Mass m_{phys}	Pole of $G^{(2)}(p)$, or zero of $\Gamma^{(2)}(p)$	Spectral decomposition	Spectrometer reading
Cross section σ	$ \mathcal{M} ^2$ from amputated $G_{\text{conn}}^{(n)}$	LSZ + $ \cdot ^2$	Event rate in detector
Decay width Γ	$\text{Im } \Sigma(m^2)$	Optical theorem	Breit–Wigner fit
Resonance mass	Real part of complex pole	Analytic continuation	Peak position in $\sigma(E)$
Coupling constant	$\Gamma^{(4)} _{\text{phys. point}}$	Renormalization condition	Fitted from σ data
Vacuum energy	$E_\Omega = \langle \Omega H \Omega \rangle$	Casimir setup	Force between plates

In every row: the left column is Fock-layer (computed by the engine), the right column is Born-layer (rendered to the observer), and the middle column is the bridge formula that connects them. The bridge always involves either the spectral decomposition, the LSZ formula, or the squared modulus—three manifestations of the Born rule applied at different levels of coarse-graining.

20 Renormalization: The Hamiltonian Side

20.1 The game engine is a finite lattice

The Fock-layer engine of a quantum field theory is a finite object. We now describe it explicitly.

The spatial lattice. Take a three-dimensional cubic lattice with N_s sites per direction and spacing a , with periodic boundary conditions (a three-torus $\mathbb{T}^3$). The spatial volume is $V = (N_s a)^3$. The allowed momenta are discrete:

$$\mathbf{k} = \frac{2\pi}{N_s a} \mathbf{n}, \quad \mathbf{n} \in \{0, 1, \dots, N_s - 1\}^3, \quad k_i \in \left[0, \frac{2\pi}{a}\right), \quad (20.1)$$

or equivalently $k_i \in (-\pi/a, \pi/a]$ (the first Brillouin zone). The number of spatial modes is N_s^3 —finite.

The Fock space. One harmonic oscillator per momentum mode $\mathbf{k}$, each truncated at occupation number $n_{\max}$ (or left untruncated for bosons—the Fock space per mode is countably infinite, but at any finite energy only finitely many modes are excited). The Fock space is:

$$\mathcal{F}_a = \bigotimes_{|\mathbf{k}| < \pi/a} \mathcal{H}_{\mathbf{k}}, \quad \dim \mathcal{F}_a \text{ is finite (with truncation) or countably infinite.} \quad (20.2)$$

For a scalar field with N_s^3 modes, each truncated at $n_{\max}$: $\dim \mathcal{F}_a = (n_{\max} + 1)^{N_s^3}$. For $N_s = 100$ and $n_{\max} = 10$: $\dim \mathcal{F}_a = 11^{10^6}$ —finite, but astronomically large. The hidden state $h_t \in \mathcal{F}_a$ is one vector in this space, of unit norm, rotating under the Hamiltonian.

Time evolution. Time is continuous (not discretised). The Schrödinger equation $\dot{h}_t = -iH(a)h_t$ evolves the Fock vector forward in time with open boundary conditions: h_0 is specified, and $h_t = e^{-iH(a)t}h_0$ for $t > 0$. The spatial lattice is a torus (periodic); the time direction is a ray (open). This is Lorentz signature: the engine rotates unitarily, preserving $\|h_t\| = 1$ at every instant.

The Hamiltonian and its parameters. The Hamiltonian on $\mathcal{F}_a$ is:

$$H(a) = \sum_{|\mathbf{k}| < \pi/a} \omega_{\mathbf{k}}(a) a_{\mathbf{k}}^\dagger a_{\mathbf{k}} + \frac{\lambda(a)}{4!} a^3 \sum_{\mathbf{n} \in \mathbb{T}^3} : \hat{\phi}(\mathbf{x}_{\mathbf{n}})^4 :, \quad (20.3)$$

with single-particle energy $\omega_{\mathbf{k}}(a) = \sqrt{(2/a)^2 \sum_i \sin^2(k_i a/2) + m^2(a)}$ (the lattice dispersion relation, reducing to $\sqrt{|\mathbf{k}|^2 + m^2(a)}$ for $|k_i|a \ll 1$), and the field operator:

$$\hat{\phi}(\mathbf{x}) = \frac{1}{\sqrt{V}} \sum_{|\mathbf{k}| < \pi/a} \frac{1}{\sqrt{2\omega_{\mathbf{k}}(a)}} [a_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{x}} + a_{\mathbf{k}}^\dagger e^{-i\mathbf{k} \cdot \mathbf{x}}]. \quad (20.4)$$

The Hamiltonian has three **engine parameters**:

$$\theta(a) = (m^2(a), \lambda(a), Z(a)). \quad (20.5)$$

These are the entries of $H(a)$ that the engine designer must specify. They depend on a because the engine's resolution determines which modes are included and how they couple.

We write $\int d^3x$ as shorthand for $a^3 \sum_{\mathbf{n}}$, and $\int d^3k/(2\pi)^3$ as shorthand for $V^{-1} \sum_{\mathbf{k}}$, for readability. Every integral in this section is a finite sum on the lattice.

What the lattice is not. The lattice is not a “regularisation” of a more fundamental continuum theory. It is not a mathematical trick that will be removed by taking $a \rightarrow 0$ at the end. The lattice IS the engine. It is a finite, well-defined, Hermitian operator on a finite-dimensional Hilbert space. The Schrödinger equation runs. The Born rule produces probabilities. The physics is complete at every finite a .

The continuum limit $a \rightarrow 0$ is a mathematical extrapolation: it asks whether the family of engines $\{H(a)\}_{a>0}$, parametrised by $\theta(a)$, produces a well-defined limit as the resolution increases. If yes, the theory has a UV completion. If no (as for the scalar ϕ^4 theory in $d = 4$, which is trivial in the continuum limit), the engine operates at finite a —and this is perfectly physical. Nature has a shortest length scale (presumably the Planck length $\ell_P \sim 10^{-35}$ m); the lattice spacing a can be identified with it, and the continuum limit is not taken.

20.2 The cinema analogy

Cinema A projects at 24 frames per second; Cinema B at 100. Both show the same scene: a ball falling with speed v . If both use the same raw displacement per frame, Cinema B plays in slow motion. The fix: adjust the displacement per frame to the frame rate, $\Delta x(\Delta t) = v \cdot \Delta t$. The displacement per frame depends on the frame rate; the perceived speed v does not.

Two cinemas	Quantum field theory
Frame rate $1/\Delta t$	Lattice resolution $1/a$
Displacement per frame Δx	Engine parameters $\theta(a)$ in $H(a)$
Perceived speed v	Born-layer observables: masses, cross sections
Δx depends on frame rate	$\theta(a)$ depends on a
Same movie in both cinemas	Same Born outputs from $H(a_1)$ and $H(a_2)$

The movie is the Born rendering. The screen is the Born layer. The engine $H(a)$ on $\mathcal{F}_a$ changes with resolution; the screen output does not. This is the entire content of renormalization, stated before any calculation.

20.3 Rendering stability: the definition

The Born-layer observables are the quantities the observer measures: the physical mass m_{phys} , the scattering cross section σ , the decay width Γ , the vacuum energy difference (Casimir force). These are computed from the engine $H(a)$ via the bridges of Section 19.7: the propagator pole for masses, the LSZ formula for cross sections, the spectral function for decay widths.

Definition. A family of engines $\{H(a)\}_{a>0}$, with parameters $\theta(a)$, satisfies **rendering stability** if all Born-layer observables are independent of the lattice spacing a to accuracy $o(a/\lambda)$:

$$\boxed{\mathcal{O}_{\text{Born}}[\theta(a)] = \mathcal{O}_{\text{phys}} + o(a/\lambda)}, \quad (20.6)$$

where λ is the observation scale—the wavelength of the probe, or equivalently, the “pixel size” of the Born-layer screen.

The correction $o(a/\lambda)$ vanishes faster than any power of a/λ as $a/\lambda \rightarrow 0$. It encodes the lattice artefacts: the discretisation errors, the finite-volume effects, the higher-dimensional operators that the lattice generates. These are suppressed when the engine resolution $1/a$ is much finer than the observation resolution $1/\lambda$:

$$a \ll \lambda \quad \Leftrightarrow \quad \text{engine resolution} \gg \text{screen resolution}. \quad (20.7)$$

The rendering-stability condition says: if the engine is fine enough, the screen output is insensitive to exactly how fine it is.

In ML terms. Rendering stability is *resolution invariance*: the performance metric (Born output) is invariant under changes of the training resolution (lattice spacing), provided the model parameters (engine parameters $\theta(a)$) are adjusted accordingly. This is exactly the property that allows a neural network trained at one resolution to transfer to another: the weights change, but the predictions don't.

The pixel size λ . What sets λ ? It is the scale at which the observer probes the system. For a scattering experiment at centre-of-mass energy E_{CM} : $\lambda \sim 1/E_{\text{CM}}$ (the de Broglie wavelength of the colliding particles). For a measurement of the hydrogen Lamb shift: $\lambda \sim 1/(E_{2S} - E_{2P}) \sim 10^{-7}$ m. For the Casimir effect at plate separation d : $\lambda \sim d$.

The observer cannot resolve structure below scale λ . The engine operates at scale $a \ll \lambda$: it contains vastly more detail than the observer can see. The rendering-stability condition says that this hidden detail washes out in the Born output.

20.4 Why do the engine parameters depend on the lattice spacing?

A natural question from ML readers: why must $\theta(a)$ depend on a at all? Why not fix m^2 , λ , Z once and for all?

The answer: because the Hamiltonian $H(a)$ on the lattice $\mathcal{F}_a$ is a *matrix*, and the matrix elements depend on how many modes are included. When the lattice is refined from spacing a to $a' < a$: new high-momentum modes (with $\pi/a' > \mathbf{k} > \pi/a$) enter the Fock space. These new modes couple to the existing ones through the interaction $\lambda\phi^4$. The coupling shifts the energy levels, the scattering amplitudes, and the vacuum structure. To maintain the same Born-layer outputs, the parameters must be adjusted to compensate.

This is not mysterious. In a neural network: adding more layers changes the representational capacity, and the weights of the existing layers must be re-tuned (via continued training) to maintain the same test-set performance. The weights depend on the architecture (lattice spacing); the performance (Born output) does not.

A concrete example: the mass parameter. At lattice spacing a , the physical mass (the pole of the propagator) is:

$$m_{\text{phys}}^2 = m^2(a) + \Sigma(m_{\text{phys}}^2; a), \quad (20.8)$$

where $\Sigma(p^2; a)$ is the self-energy—the sum of all 1PI diagrams—computed on the lattice $\mathcal{F}_a$. At one loop for ϕ^4 :

$$\Sigma(0; a) = \frac{\lambda(a)}{2} \sum_{|\mathbf{k}| < \pi/a} \frac{1}{2\omega_{\mathbf{k}}(a) V} = \frac{\lambda(a)}{2} \int_{|\mathbf{k}| < \pi/a} \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{2\omega_{\mathbf{k}}(a)}. \quad (20.9)$$

(This is a finite sum on the lattice, written as an integral for readability.) The integral scales as $1/a^2$ for $d = 4$:

$$\Sigma(0; a) = \frac{\lambda(a)}{32\pi^2} \frac{c}{a^2} + O(m^2 \ln(1/(ma))), \quad (20.10)$$

where c is a calculable numerical constant (of order 1, depending on the lattice discretisation). To hold m_{phys}^2 fixed as a changes:

$$m^2(a) = m_{\text{phys}}^2 - \Sigma(m_{\text{phys}}^2; a) \approx m_{\text{phys}}^2 - \frac{c \lambda(a)}{32\pi^2 a^2} + \dots \quad (20.11)$$

As a decreases (finer lattice): $m^2(a)$ must become more negative to compensate for the growing self-energy. The engine parameter $m^2(a)$ runs with the lattice spacing.

In the traditional language, this is called “the bare mass diverges.” In the lattice-first view: a is finite, $m^2(a)$ is a finite number of order $1/a^2$, and the fact that it scales with a is simply the statement that finer lattices have more modes that contribute to the self-energy, requiring a larger compensating mass parameter. There is no divergence and no crisis.

The coupling parameter. Similarly, the physical coupling (defined by the four-point vertex at a reference momentum) is:

$$\lambda_{\text{phys}} = \lambda(a) + \delta\lambda(a) + O(\lambda^3), \quad (20.12)$$

where $\delta\lambda(a)$ is the one-loop vertex correction (computed in the next subsection). To hold λ_{phys} fixed: $\lambda(a)$ must be adjusted as a changes.

20.5 The Feshbach formula: integrating out the short-distance shell

The Feshbach–Löwdin formula is the exact mechanism by which engine parameters change when the lattice is refined. It is a general result of linear algebra, not specific to quantum field theory: it applies to any matrix that is block-decomposed.

Setup. Fix two lattice spacings $a < a'$ (finer and coarser), with momentum cutoffs $\Lambda = \pi/a$ and $\Lambda' = \pi/a'$. The finer Fock space decomposes: $\mathcal{F}_a = \mathcal{F}_{a'} \oplus \mathcal{F}_{\text{shell}}$, where $\mathcal{F}_{\text{shell}}$ contains modes with $\Lambda' < |\mathbf{k}| < \Lambda$ (the high-momentum shell being integrated out).

Write P for the projector onto $\mathcal{F}_{a'}$ (the low-momentum subspace) and $Q = \mathbf{1} - P$ for the projector onto the shell. The Hamiltonian in block form:

$$H(a) = \begin{pmatrix} H_{PP} & H_{PQ} \\ H_{QP} & H_{QQ} \end{pmatrix}, \quad (20.13)$$

where $H_{PP} = PHP$, $H_{PQ} = PHQ$, etc.

Derivation of the Feshbach formula. The eigenvalue equation $H|\psi\rangle = E|\psi\rangle$, projected onto the P and Q subspaces:

$$H_{PP}|\psi_P\rangle + H_{PQ}|\psi_Q\rangle = E|\psi_P\rangle, \quad (20.14)$$

$$H_{QP}|\psi_P\rangle + H_{QQ}|\psi_Q\rangle = E|\psi_Q\rangle, \quad (20.15)$$

where $|\psi_P\rangle = P|\psi\rangle$ and $|\psi_Q\rangle = Q|\psi\rangle$. Solve (20.15) for $|\psi_Q\rangle$:

$$(E - H_{QQ})|\psi_Q\rangle = H_{QP}|\psi_P\rangle \quad \Rightarrow \quad |\psi_Q\rangle = \frac{1}{E - H_{QQ}} H_{QP}|\psi_P\rangle. \quad (20.16)$$

Substitute into (20.14):

$$H_{PP}|\psi_P\rangle + H_{PQ} \frac{1}{E - H_{QQ}} H_{QP}|\psi_P\rangle = E|\psi_P\rangle. \quad (20.17)$$

This is an eigenvalue equation on the P subspace alone:

$$\boxed{H_{\text{eff}}(E)|\psi_P\rangle = E|\psi_P\rangle, \quad H_{\text{eff}}(E) = H_{PP} + H_{PQ} \frac{1}{E - H_{QQ}} H_{QP}.} \quad (20.18)$$

The Feshbach effective Hamiltonian $H_{\text{eff}}(E)$ acts on the coarser Fock space $\mathcal{F}_{a'}$ and reproduces the exact eigenvalues of $H(a)$. The price: it is energy-dependent (the resolvent $1/(E - H_{QQ})$ depends on E).

The second term $H_{PQ}(E - H_{QQ})^{-1}H_{QP}$ has a clear interpretation: the low-momentum state $|\psi_P\rangle$ couples to the high-momentum shell via H_{QP} (the interaction vertex creates high-momentum modes), propagates through the shell via $(E - H_{QQ})^{-1}$ (the resolvent of the shell Hamiltonian), and returns to the low-momentum subspace via H_{PQ} (the interaction vertex destroys the high-momentum modes). This is the Fock-layer origin of “virtual processes”: they are matrix elements of the resolvent, not Born-layer events.

The energy-independent approximation. For eigenvalues E much smaller than the shell energy scale Λ' : the resolvent $1/(E - H_{QQ}) \approx -1/H_{QQ} + O(E/\Lambda'^2)$ can be expanded, giving an energy-independent effective Hamiltonian:

$$H_{\text{eff}} \approx H_{PP} - H_{PQ} \frac{1}{H_{QQ}} H_{QP} + O(E/\Lambda'). \quad (20.19)$$

This is valid when we care only about states with $E \ll \Lambda'$ —the low-energy physics. The error is $O(E/\Lambda') = O(Ea')$, which is a lattice artefact of the coarser lattice.

The effective Hamiltonian IS the coarser engine. The effective Hamiltonian H_{eff} on $\mathcal{F}_{a'}$ is the new engine at the coarser resolution a' . It has the same form as the original $H(a)$ —a polynomial in creation and annihilation operators—but with shifted parameters:

$$H_{\text{eff}} = H(a') = \sum_{|\mathbf{k}| < \pi/a'} \omega_{\mathbf{k}}(a') a_{\mathbf{k}}^\dagger a_{\mathbf{k}} + \frac{\lambda(a')}{4!} \int d^3x : \hat{\phi}^4 : + (\text{induced higher-order terms}). \quad (20.20)$$

The engine parameters have shifted: $m^2(a) \rightarrow m^2(a')$, $\lambda(a) \rightarrow \lambda(a')$. The Feshbach procedure has *derived* the dependence $\theta(a')$ on $\theta(a)$: it is the exact RG transformation. Additionally, the coarser engine may contain new interaction terms (e.g., ϕ^6 , $(\partial\phi)^4$) that were absent in the original engine—these are the “induced operators” generated by integrating out the shell.

Explicit ϕ^4 computation: the coupling shift. Consider the matrix element of H_{eff} between two low-momentum particles $|\mathbf{p}_1, \mathbf{p}_2\rangle \in \mathcal{F}_{a'}$. The second-order Feshbach correction connects this two-particle state to intermediate four-particle states (two low-momentum, two high-momentum in the shell) via two applications of the ϕ^4 vertex:

$$\delta\lambda(a') = \langle \mathbf{p}_3, \mathbf{p}_4 | H_{PQ} \frac{1}{E - H_{QQ}} H_{QP} | \mathbf{p}_1, \mathbf{p}_2 \rangle. \quad (20.21)$$

In the energy-independent approximation, with the ϕ^4 vertex providing H_{PQ} and H_{QP} , the intermediate state has two particles in the shell with momenta $\mathbf{k}$ and $\mathbf{p}_1 + \mathbf{p}_2 - \mathbf{k}$, and H_{QQ} gives the energy $2\omega_{\mathbf{k}}$. The sum over shell momenta:

$$\begin{aligned} \delta\lambda(a') &= -3\lambda(a)^2 \int_{\Lambda' < |\mathbf{k}| < \Lambda} \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{4\omega_{\mathbf{k}}^2} \\ &= -\frac{3\lambda(a)^2}{32\pi^2} \int_{\Lambda'}^{\Lambda} \frac{k^2 dk}{k^2 + m^2(a)} \\ &\approx -\frac{3\lambda(a)^2}{32\pi^2} \ln \frac{\Lambda}{\Lambda'} = -\frac{3\lambda(a)^2}{32\pi^2} \ln \frac{a'}{a}. \end{aligned} \quad (20.22)$$

(The factor of 3 counts the three Wick contractions of the two ϕ^4 vertices that produce the correct topology. The approximation $\omega_{\mathbf{k}} \approx |\mathbf{k}|$ is valid for $|\mathbf{k}| \gg m$.) The effective coupling at the coarser lattice:

$$\lambda(a') = \lambda(a) - \frac{3\lambda(a)^2}{32\pi^2} \ln \frac{a'}{a} + O(\lambda^3). \quad (20.23)$$

The coupling at the coarser lattice is smaller (in absolute value) than at the finer lattice: modes have been integrated out, and their effect has been absorbed into the shifted coupling.

Similarly, the mass parameter shifts:

$$m^2(a') = m^2(a) + \frac{\lambda(a)}{2} \int_{\Lambda' < |\mathbf{k}| < \Lambda} \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{1}{2\omega_{\mathbf{k}}(a)} = m^2(a) + \frac{c\lambda(a)}{a^2} \left(1 - \frac{a^2}{a'^2}\right) + \dots \quad (20.24)$$

The mass receives a “quadratic” correction $\propto 1/a^2$ from the shell integral. This is the statement that finer lattices have more modes contributing to the self-energy.

No “bare” and “physical” — only $\theta(a)$ and $\theta(a')$. The Feshbach procedure has related the engine parameters at two resolutions: $\theta(a)$ and $\theta(a')$. Neither is “bare”; neither is “physical.” Both are well-defined engine parameters at their respective lattice spacings. The Born-layer observables (masses, cross sections) are the same for both engines, to accuracy $O(a'/\lambda)$:

$$m_{\text{phys}}^2 = m^2(a) + \Sigma(m_{\text{phys}}^2; a) = m^2(a') + \Sigma(m_{\text{phys}}^2; a') + O(a'/\lambda). \quad (20.25)$$

The self-energy Σ changes with the lattice spacing (because the momentum sum changes), and $m^2(\cdot)$ adjusts to compensate. The physical mass m_{phys} is invariant—it is a Born-layer observable, independent of the engine’s resolution.

The beta function is the rate of change of the coupling along the family:

$$\beta(\lambda) \equiv -a \frac{d\lambda}{da} = \frac{d\lambda}{d \ln(1/a)} = \frac{3\lambda^2}{16\pi^2} + O(\lambda^3). \quad (20.26)$$

(The factor $3\lambda^2/(16\pi^2)$ is obtained from (20.22) by taking $a'/a = 1 + d\ell$ and $\ln(a'/a) \approx d\ell$, giving $d\lambda/d\ell = -3\lambda^2/(32\pi^2) \cdot (-1) \cdot 2 = 3\lambda^2/(16\pi^2)$, where the factor of 2 accounts for the Lorentzian integral $\int d^4k$ versus the spatial integral $\int d^3k$ used in the Feshbach calculation.)

20.6 The Wegner flow and the beta function

The Feshbach projection can be made continuous. The Wegner flow (1994) defines a one-parameter family of Hamiltonians $H(s)$ parametrised by a flow parameter s :

$$\frac{dH}{ds} = [[\eta(s), H(s)], H(s)], \quad \eta(s) = [H_d(s), H(s)], \quad (20.27)$$

where H_d is the diagonal part of H in the Fock basis. The flow drives $H(s)$ toward band-diagonal form: the off-diagonal matrix elements connecting states with energy difference $> \Lambda(s)$ are continuously suppressed. At each s , the Hamiltonian is unitarily equivalent to the original H : the spectrum is preserved.

The beta function is the rate of change of the coupling as the flow proceeds:

$$\beta(\lambda) = \frac{d\lambda}{d \ln(1/a)} = \frac{3\lambda^2}{16\pi^2} + O(\lambda^3). \quad (20.28)$$

This agrees with the path-integral calculation (section 23.12) and the heat-kernel extraction (section 23.12).

20.7 Renormalization of correlators

The correlators $G^{(n)}$ and $G_{\text{conn}}^{(n)}$ depend on the lattice spacing a through $\theta(a)$. The stability condition (??) applied to correlators takes the form:

The pole position is a -independent. The physical mass m_{phys} —the pole of $G^{(2)}(p; a)$ —must not depend on a . This is enforced by adjusting $m^2(a)$: the bare mass absorbs the a -dependent self-energy correction so that the pole stays fixed.

The residue requires wave-function renormalisation. The residue $Z(a) = |\langle \Omega(a) | \hat{\phi} | \mathbf{p} \rangle|^2$ of the propagator pole depends on a . Define the renormalised field $\hat{\phi}_R = \hat{\phi} / \sqrt{Z(a)}$. The renormalised propagator $G_R^{(2)}(p) = Z(a)^{-1} G^{(2)}(p; a)$ has residue 1, independent of a .

Connected correlators renormalise multiplicatively.

$$G_{\text{conn,phys}}^{(n)}(p_1, \dots, p_n) = Z(a)^{-n/2} G_{\text{conn}}^{(n)}(p_1, \dots, p_n; a) + o(a/\lambda). \quad (20.29)$$

The physical connected correlator is obtained from the bare one by multiplying by $Z^{-1/2}$ per external leg. This is the same Z that appears in the LSZ formula (section 19.3): the wave-function renormalisation connects the bare field (the engine's variable) to the physical field (the one whose correlators give a -independent Born outputs).

The 1PI vertices $\Gamma^{(n)}$ renormalise by absorbing $\theta(a)$. $\Gamma^{(2)}(p; a) = p^2 - m^2(a) - \Sigma(p; a)$. The renormalisation conditions:

$$\Gamma_R^{(2)}(p^2 = m_{\text{phys}}^2) = 0 \quad (\text{fixes } m^2(a)), \quad (20.30)$$

$$\left. \frac{d\Gamma_R^{(2)}}{dp^2} \right|_{p^2=m^2} = 1 \quad (\text{fixes } Z(a)), \quad (20.31)$$

$$\Gamma_R^{(4)}|_{\text{sym. pt}} = \lambda_{\text{phys}} \quad (\text{fixes } \lambda(a)). \quad (20.32)$$

These three conditions determine the three bare parameters $m^2(a)$, $Z(a)$, $\lambda(a)$ as functions of a , in terms of the three Born-layer observables m_{phys} , λ_{phys} , and the normalisation convention. The theory is renormalisable because three conditions suffice for all n -point functions.

20.8 What renormalisability means

A renormalisable theory: finitely many bare parameters $\theta(a)$ suffice for the stability condition (??) to hold for all Born-layer observables. The ϕ^4 theory in $d = 4$ is renormalisable: three parameters (m^2 , λ , Z) absorb all the a -dependence.

A non-renormalisable theory: infinitely many parameters are needed. This is not a sickness; it means the theory has many parameters—the full lattice Hamiltonian $H(a)$ with all its coefficients. It is still predictive at energies $E \ll 1/a$, where the higher-dimensional operators contribute corrections suppressed by $(Ea)^{d_n-4}$. Gravity is non-renormalisable in this sense: it works beautifully as an effective theory at $E \ll E_P$, and the lattice-level theory (whatever it is) would have many parameters.

The stability condition $o(a/\lambda)$ is the same as the statement that the irrelevant operators (those with mass dimension > 4) are suppressed at long wavelengths. Renormalisability is the statement that the relevant and marginal operators (dimension ≤ 4) are the only ones that matter for the Born rendering.

21 The Scalar Field: Path Integral, $Z[J]$, $W[J]$, $\Gamma[\bar{\phi}]$

The previous two sections built the Fock-layer engine entirely on the Hamiltonian side: the Fock space $\mathcal{F}_a$ on a spatial torus $\mathbb{T}^3$ with lattice spacing a , the Hamiltonian $H(a)$ with engine parameters $\theta(a) = (m^2(a), \lambda(a), Z(a))$, the correlators $G^{(n)}$, the generating-functional hierarchy $Z[J] \rightarrow W[J] \rightarrow \Gamma[\bar{\phi}]$, and the rendering-stability condition $\mathcal{O}_{\text{Born}}[\theta(a)] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$.

This section introduces the path integral as a second computational language for evaluating the same Fock-layer matrix elements. The architecture (the Fock-Born framework) is unchanged. Every answer agrees with the operator side. What the path integral provides is computational leverage.

The path-integral measure is completely well-defined on the lattice. On a spacetime lattice $\Lambda = \{0, \dots, N_\tau - 1\} \times \{0, \dots, N_s - 1\}^3$ with $N = N_\tau N_s^3$ sites, the path integral is:

$$Z[J; \theta(a), a] = \int_{\mathbb{R}^N} \prod_{n=1}^N d\phi_n \exp\left(iS_a[\phi; \theta(a)] + ia^4 \sum_n J_n \phi_n\right). \quad (21.1)$$

This is an ordinary N -dimensional integral with the product Lebesgue measure $\prod_n d\phi_n$. No infinite-dimensional measure theory, no functional-analysis subtlety, no mathematical ambiguity. The lattice action $S_a[\phi; \theta(a)]$ is a polynomial in the N real variables ϕ_n , with coefficients given by the engine parameters $\theta(a)$.

The “measure problem” of the continuum path integral—the question of how to define $\int \mathcal{D}\phi$ rigorously on an infinite-dimensional space—does not arise on the lattice. The continuum notation $\int \mathcal{D}\phi e^{iS}$ is shorthand for (21.1) in the limit $N \rightarrow \infty$.

Minkowski vs. Euclidean. The Minkowski path integral (21.1) has oscillatory weight e^{iS_M} with $|e^{iS_M}| = 1$. On a finite lattice, it is a well-defined integral (no convergence issue), but numerical evaluation is intractable due to catastrophic phase cancellation.

Under Wick rotation $t = -i\tau$ (equivalently, $\varepsilon_t \rightarrow -i\varepsilon_\tau$ on the lattice), the weight becomes e^{-S_E} with $S_E \geq 0$: real, positive, and absolutely convergent. The Euclidean path integral:

$$Z_E[J; \theta(a), a] = \int_{\mathbb{R}^N} \prod_{n=1}^N d\phi_n \exp\left(-S_{E,a}[\phi; \theta(a)] + a^4 \sum_n J_n \phi_n\right) \quad (21.2)$$

has three virtues: (i) absolute convergence (the quartic term ensures $S_E \rightarrow \infty$ at large ϕ), (ii) positive measure ($e^{-S_E} \geq 0$, so Monte Carlo sampling is applicable), and (iii) a concrete connection to statistical mechanics (the weight e^{-S_E} is a Boltzmann factor).

For perturbative calculations, both signatures give the same Feynman rules (related by $k_0 \leftrightarrow ik_E^0$). For non-perturbative calculations (lattice QCD, condensed-matter simulations), the Euclidean lattice is the workhorse. We will be explicit throughout about which signature is used.

The path-integral variables are bookkeeping eigenvalues. The integration variable ϕ_n at site n is a basis label in a decomposition of a Fock-layer matrix element, not a Born-layer measurement of the field at that spacetime point. No observer measures $\hat{\phi}(\mathbf{x}, t)$ at every lattice site. No Born rendering occurs. The “sum over all field configurations” is a matrix computation—a product of transfer matrices decomposed via completeness—exactly as in the quantum-mechanical path integral of Section 7.

21.1 From the Hamiltonian to the path integral: time slicing for a field theory

The derivation follows the same logic as the quantum-mechanical case (section 7), with one upgrade: the position x of a single particle is replaced by the field configuration $\phi(\mathbf{x})$ at all spatial points simultaneously.

The field-configuration eigenstates. In section 18, the field operator $\hat{\phi}(\mathbf{x})$ was expressed as a superposition of creation and annihilation operators (18.3). Define the *field-configuration eigenstates* $|\phi(\cdot)\rangle$ by:

$$\hat{\phi}(\mathbf{x}) |\phi(\cdot)\rangle = \phi(\mathbf{x}) |\phi(\cdot)\rangle \quad \text{for all } \mathbf{x}. \quad (21.3)$$

The eigenvalue $\phi(\mathbf{x})$ is a classical function: a real number assigned to each spatial point $\mathbf{x}$. The eigenstate $|\phi(\cdot)\rangle$ is a vector in Fock space $\mathcal{F}_a$ —a specific superposition of n -particle states for all n (in fact, a coherent-state-like superposition). On the lattice, with N_s spatial sites, the eigenvalue is a vector $\phi \in \mathbb{R}^{N_s}$: one real number per site.

These eigenstates form a resolution of identity:

$$\mathbf{1}_{\mathcal{F}_a} = \int \mathcal{D}\phi |\phi(\cdot)\rangle \langle \phi(\cdot)|, \quad \mathcal{D}\phi = \prod_{\mathbf{x} \in \text{lattice}} d\phi(\mathbf{x}). \quad (21.4)$$

The “integral over all field configurations” is, on the lattice, an ordinary finite-dimensional integral over $\mathbb{R}^{N_s}$. No infinite-dimensional measure theory is needed. The continuum notation $\mathcal{D}\phi$ is shorthand for the lattice product measure in the limit $N_s \rightarrow \infty$.

The propagator as a matrix element. Just as in quantum mechanics, the central object is the matrix element of e^{-iHT} between two field-configuration eigenstates:

$$K[\phi_f, T; \phi_i, 0] = \langle \phi_f(\cdot) | e^{-iHT} | \phi_i(\cdot) \rangle. \quad (21.5)$$

This is a number—a complex amplitude—that depends on two field configurations $\phi_i(\mathbf{x})$ (the initial configuration) and $\phi_f(\mathbf{x})$ (the final configuration), and the elapsed time T . On the lattice, it is the (ϕ_f, ϕ_i) entry of the matrix e^{-iHT} in the field-configuration basis: a matrix with continuous indices (ϕ_f and ϕ_i each range over $\mathbb{R}^{N_s}$).

Time slicing. Split e^{-iHT} into N_t factors of $e^{-iH\varepsilon}$ with $\varepsilon = T/N_t$, and insert the resolution of identity (21.4) between each pair:

$$K[\phi_f, T; \phi_i, 0] = \int \prod_{k=1}^{N_t-1} \mathcal{D}\phi_k \prod_{k=0}^{N_t-1} \langle \phi_{k+1}(\cdot) | e^{-iH\varepsilon} | \phi_k(\cdot) \rangle, \quad (21.6)$$

where $\phi_0 = \phi_i$ and $\phi_{N_t} = \phi_f$. This is matrix multiplication in the field-configuration basis: the sum over all intermediate field configurations $\phi_1, \dots, \phi_{N_t-1}$ is the field-theoretic analogue of summing over intermediate positions in quantum mechanics.

Each intermediate field configuration $\phi_k(\mathbf{x})$ is a *bookkeeping eigenvalue*: a basis label inserted to factorise a matrix product. It is not a field configuration that the system “visits” at time t_k . No observer measures $\hat{\phi}(\mathbf{x})$ at t_k . No Born rendering occurs. No Bohr trimming resets the state. The system does not pass through these configurations any more than a matrix product “passes through” its intermediate indices.

The short-time matrix element. For small ε , each factor $\langle \phi_{k+1} | e^{-iH\varepsilon} | \phi_k \rangle$ can be evaluated by expanding $H = T_{\text{kin}}[\hat{\pi}] + V[\hat{\phi}]$, where $\hat{\pi}(\mathbf{x})$ is the canonical momentum conjugate to $\hat{\phi}(\mathbf{x})$. Inserting the resolution of identity in the momentum basis $\mathbf{1} = \int \mathcal{D}\pi |\pi\rangle \langle \pi|$ and performing the Gaussian integral over π (exactly as in the single-particle case, but now at every spatial point simultaneously):

$$\langle \phi_{k+1} | e^{-iH\varepsilon} | \phi_k \rangle = \mathcal{N} \exp \left(i\varepsilon \sum_{\mathbf{x}} a^3 \left[\frac{1}{2} \left(\frac{\phi_{k+1}(\mathbf{x}) - \phi_k(\mathbf{x})}{\varepsilon} \right)^2 - \frac{1}{2} |\nabla_a \phi_k|^2 - V(\phi_k(\mathbf{x})) \right] \right), \quad (21.7)$$

where ∇_a is the lattice gradient, $V(\phi) = \frac{1}{2}m^2\phi^2 + \frac{\lambda}{4!}\phi^4$, and $\mathcal{N}$ is a normalisation constant. The exponent is $i\varepsilon a^3 \sum_{\mathbf{x}} \mathcal{L}(\phi_k, \dot{\phi}_k)$, where $\mathcal{L}$ is the Lagrangian density $\mathcal{L} = \frac{1}{2}\dot{\phi}^2 - \frac{1}{2}(\nabla\phi)^2 - V(\phi)$ and $\dot{\phi}_k \approx (\phi_{k+1} - \phi_k)/\varepsilon$.

The Lagrangian density re-emerges from the Hamiltonian by Gaussian integration over the conjugate momentum π —a mathematical step inside the Fock layer, identical in structure to the single-particle case.

The path integral formula. Substituting (21.7) into (21.6) and taking $N_t \rightarrow \infty$:

$$K[\phi_f, T; \phi_i, 0] = \int_{\phi(\cdot, 0) = \phi_i}^{\phi(\cdot, T) = \phi_f} \mathcal{D}\phi e^{iS[\phi]}, \quad S[\phi] = \int_0^T dt \int d^3\mathbf{x} \mathcal{L}(\phi, \partial_\mu\phi). \quad (21.8)$$

The “sum over all field configurations” is the continuum limit of the time-sliced product of short-time matrix elements. On the lattice (which is the physical Fock layer), the integral is over $\phi_{k,\mathbf{x}} \in \mathbb{R}$ at each of $N_t \times N_s^3$ spacetime lattice sites: a finite-dimensional integral, well-defined without any measure-theoretic complications.

Remark. The path integral in quantum field theory is often presented as a “sum over all possible field configurations,” with each configuration weighted by $e^{iS[\phi]}$. This language is seductive but misleading. A “field configuration” $\phi(\mathbf{x}, t)$ is not a sequence of Born-layer outcomes. It is a bookkeeping eigenvalue—an index in a matrix product—that runs over all values in the integral. The word “configuration” suggests the field is “in” configuration ϕ at time t , as if a measurement had been performed. But no measurement occurs at intermediate times. The field-configuration basis is one of many possible bases (the momentum basis, the coherent-state basis, the number basis) for computing the same matrix element. Different bases give different-looking integrals but the same answer. The “configurations” are computational intermediaries, not physical states.

21.2 The path integral computes Schwinger’s $Z[J]$

Schwinger’s generating functional, defined on the operator side as the vacuum-to-vacuum amplitude in the presence of a source (18.19), has a path-integral representation obtained by inserting (21.8) into the operator definition. The source $J(x)$ couples linearly to the field in the action:

$$Z[J] = \int \mathcal{D}\phi e^{i(S[\phi] + \int d^4x J(x)\phi(x))}. \quad (21.9)$$

This is the same $Z[J]$ as in (18.19): the operator definition and the path-integral representation are related by inserting complete sets of field-configuration eigenstates (21.4) and using (21.6). Schwinger’s $Z[J]$ is the *definition*; the path integral is one *method* for computing it.

On the lattice. The path integral (21.9) is, on a spacetime lattice with $N = N_t \times N_s^3$ sites:

$$Z[J; a] = \int_{\mathbb{R}^N} \prod_{n=1}^N d\phi_n \exp\left(i[S_a[\phi] + a^4 \sum_n J_n \phi_n]\right), \quad (21.10)$$

where $S_a[\phi]$ is the lattice action (a polynomial in the ϕ_n and their lattice differences) and the sum $a^4 \sum_n J_n \phi_n$ is the lattice version of $\int d^4x J\phi$. At finite a with finite N , this is an ordinary N -dimensional integral. The lattice spacing a is a parameter of the engine, not a regulator to be removed.

Correlators from $Z[J]$. The operator-side relation (18.21) carries over unchanged:

$$G^{(n)}(x_1, \dots, x_n) = \frac{1}{i^n} \frac{\delta^n Z[J]}{\delta J(x_1) \cdots \delta J(x_n)} \Big|_{J=0} = \frac{\int \mathcal{D}\phi \phi(x_1) \cdots \phi(x_n) e^{iS[\phi]}}{\int \mathcal{D}\phi e^{iS[\phi]}}. \quad (21.11)$$

The correlator is a weighted average of the product $\phi(x_1) \cdots \phi(x_n)$ over all field configurations, with weight $e^{iS[\phi]}/Z$. On the operator side, it is a vacuum expectation value of time-ordered field operators (18.5). The two expressions compute the same number—a Fock-layer matrix element—by different methods.

21.3 Worked example: the free-field path integral

For the free scalar field, $V(\phi) = \frac{1}{2}m^2\phi^2$ and the action is quadratic:

$$S_0[\phi] = \int d^4x \left[\frac{1}{2}(\partial_\mu \phi)^2 - \frac{1}{2}m^2\phi^2 \right] = -\frac{1}{2} \int d^4x \phi(x) (\partial^2 + m^2) \phi(x), \quad (21.12)$$

where the second equality uses integration by parts (and drops the boundary term). Define the Klein–Gordon operator $D(x, y) = -(\partial^2 + m^2)\delta^4(x - y)$. Then:

$$S_0[\phi] + \int J\phi = -\frac{1}{2} \int d^4x d^4y \phi(x) D(x, y) \phi(y) + \int d^4x J(x) \phi(x). \quad (21.13)$$

This is a Gaussian integral in ϕ , and Gaussian integrals can be evaluated exactly by completing the square—the same algebraic step as evaluating $\int dx e^{-\frac{1}{2}ax^2+bx} = \sqrt{2\pi/a} e^{b^2/2a}$.

Completing the square. Shift the integration variable: $\phi(x) = \chi(x) + \bar{\phi}(x)$, where $\bar{\phi}$ is chosen so that the linear term in χ vanishes. The condition is $D\bar{\phi} = J$, i.e., $\bar{\phi}(x) = \int d^4y D^{-1}(x, y) J(y)$, where D^{-1} is the inverse of the Klein–Gordon operator:

$$D^{-1}(x, y) = \int \frac{d^4p}{(2\pi)^4} \frac{i}{p^2 - m^2 + i\varepsilon} e^{-ip \cdot (x-y)} = iG_0^{(2)}(x, y). \quad (21.14)$$

After the shift:

$$S_0[\phi] + \int J\phi = -\frac{1}{2} \int \chi D \chi + \frac{1}{2} \int d^4x d^4y J(x) iG_0^{(2)}(x, y) J(y). \quad (21.15)$$

The χ integral is J -independent (it cancels in $Z_0[J]/Z_0[0]$), so:

$$\boxed{Z_0[J] = Z_0[0] \exp\left(-\frac{1}{2} \int d^4x d^4y J(x) G_0^{(2)}(x, y) J(y)\right)}. \quad (21.16)$$

This is the complete, exact generating functional for the free scalar field. It is Gaussian in J , and all correlators follow by differentiation.

The free two-point function. Differentiating twice:

$$G_0^{(2)}(x_1, x_2) = \frac{1}{i^2} \frac{\delta^2 Z_0[J]}{\delta J(x_1) \delta J(x_2)} \Big|_{J=0} = G_0^{(2)}(x_1, x_2) = \int \frac{d^4p}{(2\pi)^4} \frac{i}{p^2 - m^2 + i\varepsilon} e^{-ip \cdot (x_1 - x_2)}. \quad (21.17)$$

The path integral reproduces the free propagator obtained on the operator side from Wick’s theorem. The agreement is a consistency check, not a coincidence: the path integral was derived from the operator framework by inserting complete sets of states.

Higher correlators and Wick’s theorem. The free four-point function, obtained by differentiating (21.16) four times with respect to J :

$$G_0^{(4)}(x_1, x_2, x_3, x_4) = G_0^{(2)}(x_1, x_2) G_0^{(2)}(x_3, x_4) + G_0^{(2)}(x_1, x_3) G_0^{(2)}(x_2, x_4) + G_0^{(2)}(x_1, x_4) G_0^{(2)}(x_2, x_3). \quad (21.18)$$

This is Wick’s theorem: all free correlators factorise into products of propagators, summed over all pairings (“contractions”). In the path-integral derivation, Wick’s theorem is a consequence of the Gaussian structure of $Z_0[J]$: for a Gaussian distribution, all cumulants beyond the second vanish. On the operator side, the same result follows from the canonical commutation relations. The two proofs are different; the answer is the same.

21.4 $W[J]$ and $\Gamma[\bar{\phi}]$ from the path integral

The connected generating functional $W[J] = \ln Z[J]$ and the effective action $\Gamma[\bar{\phi}]$ (the Legendre transform of W) were defined on the operator side in section 18.5. The path integral provides explicit representations that make their physical content transparent.

$W[J]$: the connected generating functional.

$$W[J] = \ln \int \mathcal{D}\phi \, e^{i(S[\phi] + \int J\phi)}. \quad (21.19)$$

Connected correlators:

$$G_{\text{conn}}^{(n)}(x_1, \dots, x_n) = \frac{1}{i^n} \frac{\delta^n W[J]}{\delta J(x_1) \cdots \delta J(x_n)} \Big|_{J=0}. \quad (21.20)$$

For the free field, $W_0[J] = \ln Z_0[J] = \text{const} - \frac{1}{2} \int J G_0^{(2)} J$, and the only nonvanishing connected correlator is $G_{0,\text{conn}}^{(2)} = G_0^{(2)}$. All higher connected correlators vanish: a free field has no genuine multi-point interactions. This is the path-integral statement that a Gaussian distribution has no cumulants beyond the second.

The classical field. The vacuum expectation value of $\hat{\phi}(x)$ in the presence of J :

$$\bar{\phi}(x) = \frac{\delta W[J]}{\delta J(x)} = \frac{\int \mathcal{D}\phi \, \phi(x) \, e^{i(S + \int J\phi)}}{\int \mathcal{D}\phi \, e^{i(S + \int J\phi)}}. \quad (21.21)$$

This is a weighted average of the field $\phi(x)$ over all configurations, with weight $e^{iS + \int J\phi} / Z[J]$. The map $J \mapsto \bar{\phi}$ is invertible (for small J), and the inverse $J[\bar{\phi}]$ is needed for the Legendre transform.

$\Gamma[\bar{\phi}]$: the effective action from the path integral.

$$\Gamma[\bar{\phi}] = W[J] - \int d^4x \, J(x) \bar{\phi}(x), \quad (21.22)$$

with $J = J[\bar{\phi}]$. Taking the exponential of both sides and using $W = \ln Z$:

$$e^{i\Gamma[\bar{\phi}]} = \int \mathcal{D}\phi \, e^{iS[\phi] - i \int J[\bar{\phi}] \cdot (\phi - \bar{\phi})}. \quad (21.23)$$

This has an elegant interpretation: $\Gamma[\bar{\phi}]$ is the “constrained path integral”—integrate over all field configurations ϕ that fluctuate around the prescribed expectation value $\bar{\phi}$, weighted by e^{iS} . The constraint is enforced by the source $J[\bar{\phi}]$, which acts as a Lagrange multiplier.

21.5 The saddle-point (loop) expansion

The effective action $\Gamma[\bar{\phi}]$ encodes all quantum corrections to the classical action $S[\bar{\phi}]$. The loop expansion organises these corrections as a power series in $\hbar$ (which we have been setting to 1, but restoring momentarily for bookkeeping).

The split. Write $\phi = \bar{\phi} + \sqrt{\hbar}\xi$, where ξ is the fluctuation field. The action:

$$S[\bar{\phi} + \sqrt{\hbar}\xi] = S[\bar{\phi}] + \sqrt{\hbar} \int \frac{\delta S}{\delta \phi} \Big|_{\bar{\phi}} \xi + \frac{\hbar}{2} \int \xi \frac{\delta^2 S}{\delta \phi \delta \phi} \Big|_{\bar{\phi}} \xi + O(\hbar^{3/2}). \quad (21.24)$$

The linear term vanishes when $\bar{\phi}$ satisfies the classical equation of motion $\delta S/\delta \phi|_{\bar{\phi}} = -J$ (which is guaranteed by the Legendre-transform construction). Define the fluctuation operator:

$$D[\bar{\phi}](x, y) = - \frac{\delta^2 S}{\delta \phi(x) \delta \phi(y)} \Big|_{\bar{\phi}} = (\partial^2 + m^2 + \frac{\lambda}{2} \bar{\phi}^2(x)) \delta^4(x - y). \quad (21.25)$$

This is the Klein–Gordon operator with an effective $\bar{\phi}$ -dependent mass: the fluctuation ξ propagates in the background of the classical field $\bar{\phi}$.

Gaussian integration: the one-loop result. At leading order in $\hbar$ (one loop), only the quadratic term in ξ survives. The Gaussian integral over ξ :

$$\int \mathcal{D}\xi \, e^{-\frac{i\hbar}{2} \int \xi D[\bar{\phi}] \xi} = (\det D[\bar{\phi}])^{-1/2}, \quad (21.26)$$

so:

$$\Gamma[\bar{\phi}] = S[\bar{\phi}] + \frac{i}{2} \ln \det D[\bar{\phi}] + O(\hbar^2) = S[\bar{\phi}] + \Gamma^{(1)}[\bar{\phi}] + \dots \quad (21.27)$$

The one-loop correction is $\Gamma^{(1)} = \frac{i}{2} \ln \det D = \frac{i}{2} \text{Tr} \ln D$. Using the identity $\ln \det = \text{Tr} \ln$ and the proper-time (heat-kernel) representation $\ln D = - \int_0^\infty \frac{ds}{s} e^{-sD}$:

$$\Gamma^{(1)}[\bar{\phi}] = -\frac{1}{2} \int_0^\infty \frac{ds}{s} \text{Tr} e^{-sD[\bar{\phi}]}. \quad (21.28)$$

This is the same one-loop effective action computed by the heat kernel in section 23.11. The path integral provides the derivation; the heat kernel provides the evaluation.

The loop expansion summarised.

Order	Term in Γ	Physical content
Tree ($\hbar^0$)	$S[\bar{\phi}]$ (classical action)	Classical dynamics; no quantum effects
One loop ($\hbar^1$)	$\frac{i}{2} \text{Tr} \ln D[\bar{\phi}]$	Gaussian fluctuations around $\bar{\phi}$; Casimir energy; Coleman–Weinberg potential
Two loops ($\hbar^2$)	Non-Gaussian corrections	Leading interaction effects among fluctuations
n loops	$O(\hbar^n)$	n -th order in the coupling

The loop expansion is a systematic approximation in the strength of quantum fluctuations. At tree level, $\Gamma = S$: the quantum effective action reduces to the classical action, and the stationary point of Γ is the classical field equation. Each loop adds one power of $\hbar$ (or equivalently, one power of the coupling λ) and one additional layer of fluctuation.

Remark. The tree-level approximation $\Gamma \approx S$ is the regime where the Born output is near-deterministic: the Born distribution $P(\bar{\phi})$ is concentrated near the classical solution, and the variance is $O(\hbar)$. This is the classical regime discussed in the preface: the Born interface is inert because the distribution is a delta. The loop expansion is the systematic extension beyond this regime: each loop adds one layer of Born-layer randomness to the rendering.

21.6 Feynman diagrams from the perturbative expansion

The loop expansion treats the full action non-perturbatively (the background $\bar{\phi}$ can be large). The alternative is perturbation theory: treat the interaction $\frac{\lambda}{4!}\phi^4$ as a small correction to the free action S_0 . This produces Feynman diagrams.

The expansion. Write $S = S_0 + S_{\text{int}}$ with $S_{\text{int}} = -\int d^4x \frac{\lambda}{4!}\phi^4$. Then:

$$Z[J] = \int \mathcal{D}\phi e^{iS_0[\phi] + iS_{\text{int}}[\phi] + i\int J\phi} = \sum_{n=0}^{\infty} \frac{i^n}{n!} \int \mathcal{D}\phi (S_{\text{int}}[\phi])^n e^{iS_0[\phi] + i\int J\phi}. \quad (21.29)$$

Each term is a polynomial in ϕ times a Gaussian weight $e^{iS_0 + i\int J\phi}$. The Gaussian integral evaluates by Wick's theorem: contract all ϕ 's in pairs using the free propagator $G_0^{(2)}(x, y)$. The result is a sum of terms, each of which is a product of propagators integrated over internal (vertex) positions.

Feynman rules for ϕ^4 theory. Each term in the expansion corresponds to a diagram with the following rules:

Propagator: each internal line connecting points x and y contributes a factor $iG_0^{(2)}(x, y)$. In momentum space (after Fourier-transforming all positions):

$$\text{Propagator:} \quad p \text{ ————— } = \frac{i}{p^2 - m^2 + i\varepsilon}. \quad (21.30)$$

Vertex: each four-point vertex contributes $-i\lambda$, with momentum conservation $(2\pi)^4\delta^4(\sum_i p_i)$ at each vertex:

$$\text{Vertex:} \quad \begin{array}{c} \diagup \quad \diagdown \\ \bullet \\ \diagdown \quad \diagup \end{array} = -i\lambda. \quad (21.31)$$

Loop momenta: integrate $\int d^4k/(2\pi)^4$ over each undetermined internal momentum (one per independent loop).

Symmetry factor: divide by the symmetry factor of the diagram—the number of permutations of internal lines and vertices that leave the diagram unchanged.

External legs: for each external point x_i , attach a propagator and (for S -matrix elements) apply the LSZ formula.

The first correction to the propagator. At order λ (one vertex, one loop), the self-energy diagram is:

$$-i\Sigma^{(1)}(p) = p \text{ ————— } \begin{array}{c} \circ \\ | \\ \bullet \end{array} \text{ ————— } p = \frac{(-i\lambda)}{2} \int \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2 + i\varepsilon}. \quad (21.32)$$

The factor $1/2$ is the symmetry factor. The integral diverges: in $d = 4$, it scales as $\int dk k^3/k^2 = \int dk k \rightarrow \infty$ (a quadratic divergence).

On the lattice, the integral is cut off at $|\mathbf{k}| < \pi/a$:

$$-i\Sigma^{(1)}(p; a) = \frac{(-i\lambda(a))}{2} \int_{|\mathbf{k}| < \pi/a} \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2(a) + i\varepsilon} = \frac{\lambda(a)}{32\pi^2} \left[\frac{\pi^2}{a^2} - m^2(a) \ln \frac{\pi/a}{m(a)} + \dots \right]. \quad (21.33)$$

The quadratic term π^2/a^2 depends on the lattice spacing. The rendering-stability condition (section ??) requires that the physical mass $m_{\text{phys}}^2 = m^2(a) + \text{Re} \Sigma(m_{\text{phys}}^2; a)$ does *not* depend on a . This is enforced by choosing $m^2(a)$ to absorb the a -dependent piece: $m^2(a) = m_{\text{phys}}^2 - \Sigma(m_{\text{phys}}^2; a) + O(\lambda^2)$. The bare mass runs with the lattice spacing; the physical mass does not. This is mass renormalization.

The first correction to the vertex. At order λ^2 (two vertices, one loop), the vertex correction:

$$\delta\Gamma^{(4)}(p_1, p_2, p_3, p_4) = \frac{(-i\lambda(a))^2}{2} \int_{|\mathbf{k}| < \pi/a} \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2} \frac{i}{(p_1 + p_2 - k)^2 - m^2} + (2 \text{ channels}). \quad (21.34)$$

The integral diverges logarithmically: $\int dk k^3/k^4 = \int dk/k \rightarrow \ln(\pi/a)$. The rendering-stability condition absorbs the logarithm into $\lambda(a)$:

$$\lambda(a) = \lambda_{\text{phys}} + \frac{3\lambda_{\text{phys}}^2}{16\pi^2} \ln \frac{\pi/a}{\mu} + O(\lambda^3), \quad (21.35)$$

where μ is the momentum scale at which λ_{phys} is defined (the renormalization point). The bare coupling runs logarithmically with the lattice spacing. The physical coupling, defined at a fixed momentum scale, does not.

What the diagrams are—and what they are not. Each Feynman diagram is a Fock-layer bookkeeping device: it organises the terms in the Taylor expansion of a Fock-layer matrix element $G_{\text{conn}}^{(n)}$. The vertices do not depict events (particles emitted, absorbed, scattered). They depict the algebraic structure of the perturbative expansion of the matrix element.

The temptation to narrate diagrams in Born-layer language is strong: “the electron emits a photon, which is absorbed by another electron.” This is a conflation. The diagram computes a term in a matrix element. The “photon” is a propagator factor $G_0^{(2)}(k)$ —a bookkeeping Green’s function, not a Born-layer event. No observer detects the “exchanged photon.” No Born rendering occurs at the vertex. The only Born-layer observable is the final cross section $\sigma \propto |\sum_{\text{diagrams}} \mathcal{M}|^2$, obtained after summing all diagrams, amputating external legs, and squaring.

The diagrams are identical to those from the operator-side Dyson series (section 18.7). The path integral and the operator formalism produce the same combinatorics. This is not a coincidence: the path integral was derived from the operator framework by inserting complete sets of states. The shared language of Feynman diagrams does not change the fact that both methods are computing the same Fock-layer matrix elements.

21.7 The $i\varepsilon$ prescription

The Minkowski path integral $\int \mathcal{D}\phi e^{iS_M}$ is oscillatory: the integrand has unit modulus $|e^{iS_M}| = 1$, and the integral converges (if it converges at all) only through cancellations between rapidly oscillating phases. The $i\varepsilon$ prescription $m^2 \rightarrow m^2 - i\varepsilon$ (equivalently $T \rightarrow T(1 - i\varepsilon)$) tilts the time contour slightly into the complex plane.

What $i\varepsilon$ does. The prescription performs two functions simultaneously:

Convergence. The integral acquires a damping factor $e^{-\varepsilon \int \phi^2}$ that suppresses large-field configurations. Without this damping, the Minkowski path integral is at best conditionally convergent (it converges only if the integration contour is taken in a specific order). With the damping, the integral is absolutely convergent.

Vacuum selection. In the operator framework, the $i\varepsilon$ implements the Gell-Mann–Low projection (18.47): it selects the interacting vacuum $|\Omega\rangle$ by killing excited states at large times. The factor $e^{-iHT(1-i\varepsilon)} = e^{-iHT} \cdot e^{-\varepsilon HT}$ damps all states exponentially in T , with the ground state damped least (it has the smallest eigenvalue of H). In the limit $T \rightarrow \infty$, only the ground state survives.

$i\varepsilon$ on the lattice. On a lattice with finite N sites, the Minkowski path integral (21.10) is a finite-dimensional oscillatory integral. Such integrals are well-defined (no convergence issues), but they are numerically intractable because the integrand oscillates wildly. The $i\varepsilon$ prescription is not needed for mathematical well-definedness on a finite lattice; it is needed for selecting the vacuum state and for making the connection to the Euclidean formulation (section 23).

In the Euclidean formulation ($t \rightarrow -i\tau$), the $i\varepsilon$ is unnecessary because e^{-S_E} is already a positive, real, damped integrand: absolutely convergent and ground-state-projecting without any additional prescription. The Euclidean path integral is the natural computational tool; the Minkowski path integral is related to it by analytic continuation.

21.8 The bridges to Born-layer observables

The path integral computes Fock-layer quantities: $Z[J]$, $W[J]$, $\Gamma[\bar{\phi}]$, and their functional derivatives ($G^{(n)}$, $G_{\text{conn}}^{(n)}$, $\Gamma^{(n)}$). None of these are Born-layer observables. They are internal properties of the engine—amplitudes, matrix elements, vertex functions—that the engine computes but the observer never sees directly.

The bridges from Fock-layer quantities to Born-layer observables are the same as on the Hamiltonian side (section 19), because the two computational languages produce the same Fock-layer quantities. We catalogue them here for completeness, now with the path-integral expressions.

Bridge 1: expectation values. The Born expected value of an observable $\hat{\mathcal{O}}$ in the vacuum:

$$\langle \hat{\mathcal{O}} \rangle_{\text{Born}} = \langle \Omega | \hat{\mathcal{O}} | \Omega \rangle = \frac{\int \mathcal{D}\phi \mathcal{O}[\phi] e^{iS[\phi]}}{\int \mathcal{D}\phi e^{iS[\phi]}}. \quad (21.36)$$

For $\hat{\mathcal{O}} = \hat{\phi}(x)$: this is $\bar{\phi}(x)|_{J=0} = \delta W / \delta J|_{J=0}$. For a field with unbroken symmetry, $\bar{\phi}|_{J=0} = 0$. For a field with spontaneous symmetry breaking, $\bar{\phi}|_{J=0} = v \neq 0$: the vacuum has a nonzero field expectation value. The VEV v is a Born-layer observable (it determines the masses of the gauge bosons through the Higgs mechanism).

Bridge 2: masses from poles. The Fourier-transformed propagator $\tilde{G}_{\text{conn}}^{(2)}(p)$ has a pole at $p^2 = m_{\text{phys}}^2$. The mass is extracted from the pole position:

$$m_{\text{phys}}^2 = \text{zero of } \Gamma^{(2)}(p), \quad \text{where } \Gamma^{(2)}(p) = p^2 - m^2(a) - \Sigma(p; a). \quad (21.37)$$

The mass is a Born-layer observable—what the spectrometer reports—computed from a Fock-layer quantity (the pole of the propagator).

Bridge 3: cross sections via LSZ. The LSZ formula (19.3) extracts $\langle f | S | i \rangle$ from $G_{\text{conn}}^{(n)}$. The cross section $\sigma = |\langle f | S | i \rangle|^2$ is a Born probability. The path integral provides $G_{\text{conn}}^{(n)}$ via Feynman diagrams; LSZ converts it to an S -matrix element; the squared modulus converts it to a Born-layer event rate.

Bridge 4: decay widths. $\Gamma_{\text{decay}} = -\text{Im} \Sigma(m^2)/m$, where Σ is the self-energy extracted from $\Gamma^{(2)}$. The decay width is Born-layer (the rate at which detector clicks accumulate); the self-energy Σ is Fock-layer (a functional derivative of Γ).

Bridge 5: thermodynamics via the Euclidean path integral. The thermal partition function:

$$\mathcal{Z}(\beta) = \text{Tr} e^{-\beta H} = \int \mathcal{D}\phi e^{-S_E[\phi]}, \quad (21.38)$$

with periodic boundary conditions $\phi(\mathbf{x}, 0) = \phi(\mathbf{x}, \beta)$ in the Euclidean time direction. This is the Euclidean path integral on a spacetime with the topology of a cylinder: the Euclidean time direction is compactified to a circle of circumference $\beta = 1/T$.

The free energy $F = -T \ln \mathcal{Z}$, the pressure $P = -\partial F / \partial V$, the entropy $S = -\partial F / \partial T$, and the specific heat $c_V = T \partial S / \partial T$ are all Born-layer thermodynamic observables: what a thermometer, a pressure gauge, and a calorimeter report.

Bridge	Fock-layer intermediary	Born-layer observable
Expectation value	$\delta W / \delta J$	$\langle \hat{\mathcal{O}} \rangle$ (VEV, condensate)
Mass	Pole of $G_{\text{conn}}^{(2)}(p)$	m_{phys} from spectrometer
Cross section	$G_{\text{conn}}^{(n)}$ via LSZ	$\sigma = \mathcal{M} ^2 \times (\text{phase space})$
Decay width	$\text{Im} \Sigma(m^2)$	Γ from Breit–Wigner fit
Thermodynamics	$\ln \mathcal{Z}(\beta)$ from Euclidean PI	F, P, s, c_V

Every row has the same structure: the path integral computes a Fock-layer quantity (left), and a bridge formula converts it to a Born-layer observable (right). The bridge always involves either the spectral decomposition, the LSZ formula, or the squared modulus—three manifestations of the Born rule applied at different levels of coarse-graining.

21.9 The rendering-stability workflow: from $\theta(a)$ to $o(a/\lambda)$

We now make the renormalization procedure fully explicit, in path-integral language, using the rendering-stability framework of Section 20.3. The workflow has three steps.

Step 1: compute the Born-layer observable $\mathcal{O}(a)$ as a function of $\theta(a)$. Fix a lattice spacing a and engine parameters $\theta(a) = (m^2(a), \lambda(a), Z(a))$. Compute the desired Born-layer observable—a mass, a cross section, a decay width—from the Fock-layer correlators via the bridges.

For the physical mass: compute the propagator’s self-energy at one loop:

$$\Sigma(p^2; \theta(a), a) = \frac{\lambda(a)}{2} \int_{|k| < \pi/a} \frac{d^4 k}{(2\pi)^4} \frac{i}{k^2 - m^2(a) + i\varepsilon} + O(\lambda^2). \quad (21.39)$$

This is a finite integral on the lattice, and it is a function of both $\theta(a)$ (through $m^2(a)$ and $\lambda(a)$ in the integrand) and a (through the cutoff π/a). The physical mass is:

$$m_{\text{phys}}^2[\theta(a), a] = m^2(a) + \text{Re} \Sigma(m_{\text{phys}}^2; \theta(a), a). \quad (21.40)$$

At this stage, m_{phys} depends on both $\theta(a)$ and a . This is the “raw output” of the engine at resolution a .

For the scattering cross section: compute the vertex correction $\delta\Gamma^{(4)}$ and the full amplitude $\mathcal{M}$ via Feynman diagrams on the lattice. The cross section $\sigma[\theta(a), a] = |\mathcal{M}[\theta(a), a]|^2 \times (\text{phase space})$ depends on $\theta(a)$ and a .

Step 2: choose $\theta(a)$ so that $\mathcal{O}(a)$ is a -independent to $o(a/\lambda)$. The rendering-stability condition demands:

$$\mathcal{O}_{\text{Born}}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda). \quad (21.41)$$

This is a system of equations that determines $\theta(a)$ as a function of a and the physical observables. For ϕ^4 theory at one loop, the three rendering conditions are:

$$m^2(a) + \text{Re } \Sigma(m_{\text{phys}}^2; a) = m_{\text{phys}}^2 + o(a/\lambda), \quad (21.42)$$

$$1 + \frac{\partial}{\partial p^2} \text{Re } \Sigma(p^2; a) \Big|_{p^2=m_{\text{phys}}^2} = Z(a)^{-1} + o(a/\lambda), \quad (21.43)$$

$$\lambda(a) + \delta\Gamma^{(4)}(s_0, t_0, u_0; a) = \lambda_{\text{phys}} + o(a/\lambda), \quad (21.44)$$

where (s_0, t_0, u_0) is the reference momentum configuration at which the physical coupling is defined. These three equations determine the three engine parameters:

$$m^2(a) = m_{\text{phys}}^2 - \text{Re } \Sigma(m_{\text{phys}}^2; a) + o(a/\lambda), \quad (21.45)$$

$$Z(a) = \left[1 + \partial_{p^2} \text{Re } \Sigma \Big|_{m^2} \right]^{-1} + o(a/\lambda), \quad (21.46)$$

$$\lambda(a) = \lambda_{\text{phys}} - \delta\Gamma^{(4)}(s_0, t_0, u_0; a) + o(a/\lambda). \quad (21.47)$$

Each engine parameter absorbs the a -dependent loop corrections so that the Born-layer output is a -independent. The engine parameters $\theta(a)$ run with the lattice spacing; the Born-layer observables ($m_{\text{phys}}, \lambda_{\text{phys}}, Z_{\text{phys}} = 1$) do not.

Step 3: verify that all other Born-layer observables are automatically a -independent.

Having fixed $\theta(a)$ by three conditions, check that every other correlator and cross section is also rendering-stable. This is the content of *renormalisability*: the ϕ^4 theory has only three types of a -dependent divergences (quadratic in the self-energy, logarithmic in the self-energy, and logarithmic in the vertex correction), and these are absorbed by the three parameters $m^2(a)$, $Z(a)$, $\lambda(a)$. Higher-point functions $G_{\text{conn}}^{(n)}$ for $n \geq 6$ are automatically finite (in $d = 4$) and a -independent at leading order.

If the theory required infinitely many conditions to achieve rendering stability (i.e., if new a -dependent structures appeared at every loop order that could not be absorbed into the existing parameters), the theory would be *non-renormalisable*: it would require infinitely many engine parameters $\theta(a)$ and would need infinitely many Born-layer measurements to fix them. Gravity is non-renormalisable in this sense: it is perfectly well-defined at any finite a , but the number of parameters needed grows with each loop order.

The workflow summarised.

Step	Input	Operation	Output
1	$\theta(a)$, lattice spacing a	Compute $\mathcal{O}[\theta(a), a]$ via Feynman diagrams on the lattice	Born observable as function of $\theta(a)$ and a
2	Measured values of 3 observables	Solve rendering conditions for $\theta(a)$	Engine parameters $m^2(a)$, $\lambda(a)$, $Z(a)$
3	$\theta(a)$ from Step 2	Compute all other $\mathcal{O}[\theta(a), a]$	Predictions: all are a -independent to $o(a/\lambda)$

This is the complete renormalization procedure. It is a system of equations, not a mystery. The “infinities” of traditional textbooks are the statement that $\theta(a)$ depends on a ; the “cancellation of infinities” is the statement that the Born-layer observables do not. On the lattice, everything is finite: $\theta(a)$ is a finite number, the integrals are finite sums, and the Born outputs are finite predictions. The word “infinity” never appears.

21.10 Wick rotation on the lattice: from Lorentz to Euclidean

The Minkowski path integral (??) has an oscillatory weight e^{iS_M} with $|e^{iS_M}| = 1$. On a finite lattice, this integral is well-defined but numerically intractable: the oscillations produce catastrophic cancellation in any Monte Carlo estimator.

The Wick rotation $t = -i\tau$ (or equivalently, rotating the time lattice spacing $\varepsilon \rightarrow -i\varepsilon_E$) converts the oscillatory weight to a damped weight:

$$e^{iS_M[\phi]} \xrightarrow{t \rightarrow -i\tau} e^{-S_E[\phi_E]}, \quad (21.48)$$

where the Euclidean action:

$$S_E[\phi_E] = a^4 \sum_{n \in \Lambda_E} \left[\frac{1}{2} \sum_{\mu} (\nabla_{\mu}^a \phi_E)_n^2 + \frac{1}{2} m^2(a) \phi_{E,n}^2 + \frac{\lambda(a)}{4!} \phi_{E,n}^4 \right] \quad (21.49)$$

is a sum over a *Euclidean* spacetime lattice $\Lambda_E = \{0, 1, \dots, N_{\tau} - 1\} \times \{0, 1, \dots, N_s - 1\}^3$ with N_{τ} sites in the (Euclidean) time direction and N_s^3 spatial sites, and $(\nabla_{\mu}^a \phi)_n = (\phi_{n+\hat{\mu}} - \phi_n)/a$ is the lattice forward difference. All four directions are now on equal footing: the Euclidean lattice is a four-dimensional hypercubic lattice with $O(4)$ symmetry (the Euclidean analogue of Lorentz invariance).

The Euclidean path integral on the lattice.

$$Z_E[J; a] = \int_{\mathbb{R}^N} \prod_{n=1}^N d\phi_{E,n} e^{-S_E[\phi_E] + a^4 \sum_n J_n \phi_{E,n}}, \quad (21.50)$$

where $N = N_{\tau} \times N_s^3$ is the total number of lattice sites. This integral has three properties that make it the workhorse of non-perturbative QFT:

- (i) *Absolutely convergent.* For $\lambda(a) > 0$, the quartic term $\phi^4/4!$ ensures $S_E \rightarrow +\infty$ as $|\phi| \rightarrow \infty$, so $e^{-S_E} \rightarrow 0$. The integral converges without any $i\varepsilon$ prescription.
- (ii) *Positive measure.* $e^{-S_E} \geq 0$ for all ϕ_E . The weight e^{-S_E}/Z_E is a probability distribution on $\mathbb{R}^N$ —a well-defined measure from which configurations can be sampled.
- (iii) *Finite-dimensional.* The integral is over $\mathbb{R}^N$ with $N = N_{\tau} N_s^3$ finite. No functional-analysis subtlety.

Computing $\mathcal{O}(a)$ in the Euclidean framework. Euclidean correlators are computed as expectation values:

$$G_E^{(n)}(x_1, \dots, x_n; \theta(a), a) = \frac{\int \prod_m d\phi_{E,m} \phi_E(x_1) \cdots \phi_E(x_n) e^{-S_E[\phi_E; \theta(a)]}}{\int \prod_m d\phi_{E,m} e^{-S_E[\phi_E; \theta(a)]}}. \quad (21.51)$$

These are ordinary expectation values in a positive probability distribution—the same mathematical object as a thermal average in statistical mechanics, and computable by Monte Carlo sampling.

The Born-layer observables are extracted from the Euclidean correlators by the same bridges as on the Minkowski side, after analytic continuation:

Masses: the Euclidean two-point function $G_E^{(2)}(\tau) = \langle \phi_E(\tau) \phi_E(0) \rangle_E$ decays as $e^{-m_{\text{phys}} \tau}$ at large Euclidean time τ . The mass is the exponential decay rate—the same quantity as the Minkowski propagator pole, related by $p^0 = ip_E^0$.

Cross sections: Euclidean n -point functions are analytically continued to Minkowski momenta ($p_E^0 \rightarrow ip^0$) and fed into the LSZ formula. The cross section $\sigma = |\mathcal{M}|^2$ is the Born-layer output.

Thermodynamics: with periodic boundary conditions in Euclidean time ($\phi_E(\mathbf{x}, 0) = \phi_E(\mathbf{x}, \beta)$), the Euclidean path integral IS the thermal partition function $\mathcal{Z}(\beta) = \text{Tr } e^{-\beta H}$, and thermodynamic observables follow by differentiation.

Rendering stability in the Euclidean framework. The rendering-stability condition (21.41) applies identically in the Euclidean framework. The workflow of Section 21.9 carries over:

Step 1: compute $G_E^{(n)}(\theta(a), a)$ on the Euclidean lattice (perturbatively or by Monte Carlo).

Step 2: extract Born-layer observables (masses from decay rates, couplings from vertex functions).

Step 3: choose $\theta(a)$ so that the Born outputs are a -independent to $o(a/\lambda)$.

The Euclidean lattice is where non-perturbative calculations are actually performed: lattice QCD, lattice Higgs models, lattice condensed-matter simulations. The Minkowski results are recovered by analytic continuation.

Hamiltonian–path-integral correspondence. The two computational languages—Hamiltonian (operator) and path-integral—compute the same Fock-layer quantities. Their correspondence is exact:

Hamiltonian (operator) side	Path integral side
Fock space $\mathcal{F}_a$, $\dim \sim (n_{\max} + 1)^{N_s^3}$	Integration domain $\mathbb{R}^{N_t \times N_s^3}$
Hamiltonian $H(a)$: Hermitian matrix	Action $S_a[\phi]$: real function of N variables
$\theta(a) = (m^2(a), \lambda(a), Z(a))$ in $H(a)$	Same $\theta(a)$ in $S_a[\phi]$
$e^{-iH(a)t}$: unitary rotation (Lorentz)	e^{iS_M} : oscillatory weight (Minkowski PI)
$e^{-H(a)\tau}$: contraction (Euclidean)	e^{-S_E} : positive weight (Euclidean PI)
Schwinger's $Z[J]$ (operator def.)	$Z[J] = \int \mathcal{D}\phi e^{i(S + \int J\phi)}$
Feshbach: integrate out shell of modes	Wilson RG: integrate out shell of momenta
Rendering stability: $\mathcal{O}[\theta(a)] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$	Same condition, same $\theta(a)$, same observables

The Hamiltonian side works with vectors and matrices in Fock space. The path-integral side works with integrals over field configurations. Both are Fock-layer computations. Both produce the same correlators, the same $Z[J]$, the same $\Gamma[\bar{\phi}]$, and the same Born-layer observables. The choice between them is a matter of computational convenience, not of physics.

The Hamiltonian side is the primary formulation in this book: $H(a)$ defines the engine, $\dot{h}_t = -iHh_t$ runs it, and the Born rule renders the output. The path integral is a derived tool: it computes specific matrix elements of e^{-iHT} by decomposing them into multi-dimensional integrals. Neither side is more fundamental; both compute the same thing; the Hamiltonian side is where the architecture (Fock-Born) is defined.

21.11 The two languages, side by side

The Hamiltonian side (sections 18–20) and the path-integral side (this section) are two computational languages for the same Fock-layer engine. Neither is more fundamental; neither produces physics that the other cannot. The following table records the dictionary:

Hamiltonian language	Path-integral language
Fock space $\mathcal{F}_a$	Space of field configurations $\phi(\mathbf{x})$
Hamiltonian $H(a)$ on $\mathcal{F}_a$	Action $S_a[\phi]$ on lattice
e^{-iHT} : unitary time evolution	$\int \mathcal{D}\phi e^{iS}$: sum over configurations
$\langle \Omega T \{ \hat{\phi}(x_1) \cdots \} \Omega \rangle$	$\int \mathcal{D}\phi \phi(x_1) \cdots e^{iS} / Z$
Wick's theorem from $[\hat{a}, \hat{a}^\dagger] = 1$	Wick's theorem from Gaussianity of e^{iS_0}
Dyson series: Taylor-expand $T \exp(-i \int H_I)$	Expand $e^{iS_{\text{int}}}$ in powers of λ
Feshbach–Löwdin: block-diagonalise H	Wilsonian RG: integrate out shell in $\int \mathcal{D}\phi$
Wegner flow $dH/ds = [[\eta, H], H]$	Polchinski or Wetterinck exact RG
Spectral gap = m_{phys}	Pole of $\Gamma^{(2)}(p)^{-1}$
Gell-Mann–Low vacuum projection	$i\varepsilon$ prescription
Same Born-layer observables	Same Born-layer observables

The Born layer does not know or care which language the engine speaks. It renders the same observables either way. The choice between the two is a choice of computational convenience, not of physics:

Use the Hamiltonian when the operator structure is transparent (the Fock-Born architecture, the Feshbach projection, the heat kernel), when the theory has a sign problem (the Minkowski path integral oscillates; the Hamiltonian formulation does not), or when the question is about the spectrum (eigenvalues of H).

Use the path integral when Lorentz invariance must be manifest (the Euclidean action is $SO(4)$ -invariant; the Hamiltonian formulation breaks this by singling out time), when gauge-fixing is needed (the Faddeev–Popov procedure is natural in path-integral language), or when Monte Carlo sampling is available (in the Euclidean formulation, $e^{-S_E} \geq 0$ is a positive weight).

Both are Fock-layer tools. Both serve the same Born-layer predictions.

21.12 Common conflations, revisited

The path integral invites several additional conflations beyond those catalogued in the preface. We list them here with their corrections.

Conflation: “The field fluctuates in the vacuum.”

The vacuum state $|\Omega\rangle$ is the ground state of $H(a)$: a single, quiet, stationary vector in $\mathcal{F}_a$. It does not fluctuate. The expectation value $\langle \hat{\phi}^2(x) \rangle_\Omega$ is nonzero (it equals the propagator at coincident points, $G^{(2)}(x, x)$), but this is a property of the state, not a process happening inside it. The phrase “vacuum fluctuations” borrows from the path integral, where the integral $\int \mathcal{D}\phi \phi^2(x) e^{iS} / Z$ is an average over many configurations—but the configurations are bookkeeping variables, not events. The vacuum is one vector. It sits still. Its $\langle \hat{\phi}^2 \rangle$ is nonzero for the same reason that a Gaussian random variable has nonzero variance: the distribution has width, even though no “fluctuation” is happening at any moment.

Conflation: “Virtual particles mediate forces.”

The propagator $G_0^{(2)}(x, y)$ in a Feynman diagram is often called a “virtual particle” when it appears as an internal line. But the propagator is a Green’s function—a mathematical factor

in the perturbative expansion of a matrix element. It has no on-shell momentum ($p^2 \neq m^2$), no definite energy-momentum, and no Born-layer detectability. Calling it a “particle” imports Born-layer language (“particle” = “something detected”) into a Fock-layer computation. The force between two charges is the Born-layer observable; the internal propagator is the Fock-layer mechanism by which the engine produces it. The mechanism is not an event in the engine; it is a term in a computation.

Conflation: “The system explores all configurations.”

The path integral $\int \mathcal{D}\phi e^{iS[\phi]}$ is sometimes interpreted as the system “exploring” or “sampling” all field configurations, with each configuration weighted by e^{iS} . But the Fock layer is one vector, rotating under e^{-iHT} . The “sum over configurations” is a computational decomposition of one matrix element into intermediate steps. The system does not explore configurations any more than a matrix product “explores” its intermediate indices. The path integral is a method of computation, not a narrative of the system’s journey.

These conflations are not harmless. They invite questions that have no answers within the formalism (“which path did the system take?”, “how many virtual particles are in the vacuum?”, “what is the field doing between measurements?”) and obscure the clean separation between the Fock layer (one vector, rotating) and the Born layer (eigenvalues, rendered to the observer).

22 Renormalization: The Path Integral Side

Section 20 developed renormalization on the Hamiltonian side: the engine $H(a)$ on the Fock space $\mathcal{F}_a$ has engine parameters $\theta(a) = (m^2(a), \lambda(a), Z(a))$ that depend on the lattice spacing a , and the rendering-stability condition requires all Born-layer observables to satisfy $\mathcal{O}_{\text{Born}}[\theta(a)] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$.

This section develops the same physics in path-integral language. The Euclidean generating functional $W_E[J; \theta(a), a] = \ln Z_E[J; \theta(a), a]$ —computed from the Euclidean path integral on a lattice—must be stable under changes of a . The two treatments are not alternatives; they are the same rendering-stability condition expressed in two computational languages.

The workflow is explicit:

1. **Compute** Born-layer observables $\mathcal{O}[\theta(a), a]$ from the lattice path integral, as functions of the engine parameters $\theta(a)$ and the lattice spacing a .
2. **Impose** the rendering-stability condition: choose $\theta(a)$ so that each $\mathcal{O}$ is a -independent to $o(a/\lambda)$.
3. **Verify** that all other Born-layer observables are automatically a -independent (renormalisability).

We proceed in five parts: the CNN analogy, the lattice generating functional with $\theta(a)$, the one-loop effective potential, the rendering conditions on $\Gamma^{(n)}$, and the beta function. Throughout, we use “engine parameters $\theta(a)$ ” rather than “bare parameters”: the word “bare” suggests something incomplete, as if the parameters needed “dressing” to become physical. They do not. They are the complete, well-defined parameters of the engine at resolution a .

22.1 The CNN binary classifier as a field theory

Before writing a single integral, we illustrate the renormalization logic with an analogy that requires no physics background: a convolutional neural network (CNN) classifying images at different resolutions.

Consider a CNN that classifies images as “cat” or “not cat.” Each image is a field: a function $\phi(x_i)$ that assigns a pixel value to each grid point x_i . The grid has spacing a (pixel size) and resolution $\Lambda = \pi/a$ (pixels per unit length).

The CNN computes a logit (log-odds score) from the image:

$$\ell[\phi; \Lambda, \theta(\Lambda)] = a^d \sum_{\text{pixels}} \sum_i \theta_i(\Lambda) \mathcal{O}_i(\phi, \partial\phi), \quad (22.1)$$

where $\mathcal{O}_i$ are local operators built from pixel values and finite differences (edges, gradients, curvatures, textures), and $\theta_i(\Lambda)$ are the learned weights at resolution Λ . The classification probability is $P = \sigma(\ell) = 1/(1 + e^{-\ell})$.

Now double the resolution: $\Lambda \rightarrow 2\Lambda$ (halve the pixel size). The grid has 2^d times more pixels. If the weights θ_i are kept fixed, the logit scales as $\ell \propto (2\Lambda)^d$: the sum over pixels has 2^d more terms, each contributing the same amount. The logit blows up, the sigmoid saturates, and the classifier becomes useless—it classifies *everything* as “cat” (or everything as “not cat”).

This is the field-theory analogue of the divergent self-energy: holding the bare parameters fixed while changing the resolution produces an unphysical result. The Born-layer output (the classification probability P) depends wildly on the engine’s resolution.

The fix is to *rescale the weights with the resolution* so that $\ell = O(1)$ at any resolution. Different operators require different rescalings, determined by their *mass dimension*—the number of

powers of the resolution that the operator contributes to the logit:

Operator	$[\mathcal{O}]$	θ scaling	CNN analogue	Class
1	0	Λ^{-4}	Bias	Relevant
ϕ^2	2	Λ^{-2}	Contrast	Relevant
$(\partial\phi)^2$	4	$\ln \Lambda$	Edge filter	Marginal
ϕ^4	4	$\ln \Lambda$	Quartic	Marginal
ϕ^6	6	Λ^2	6th-order	Irrelevant

The classification is simple:

Relevant operators ($[\mathcal{O}] < 4$ in $d = 4$): their weights grow as $\Lambda \rightarrow \infty$ —they dominate the logit at high resolution. They must be tuned carefully. The mass term $m^2\phi^2$ is relevant: the bare mass $m^2(a)$ must be adjusted by $O(1/a^2)$ as a changes.

Marginal operators ($[\mathcal{O}] = 4$): their weights grow logarithmically. They run slowly and require fine-tuning only at the logarithmic level. The quartic coupling $\lambda\phi^4$ is marginal: $\lambda(a)$ runs as $\ln(1/a)$.

Irrelevant operators ($[\mathcal{O}] > 4$): their weights shrink as $\Lambda \rightarrow \infty$ —they wash out at coarse resolution. The fine-grained pixel-level detail (the 6th-order texture correlations, the 8th-order pattern interactions) vanishes when the image is downsampled. Irrelevant operators contribute corrections suppressed by $(a/\lambda)^{[\mathcal{O}]-4}$, which vanish in the rendering-stability condition.

This is Wilson’s operator classification, visible without any physics: relevant and marginal operators survive resolution changes; irrelevant operators do not. The CNN’s learned weights $\theta_i(\Lambda)$ are the analogue of the bare parameters $\theta(a)$ in the QFT Hamiltonian. The classification probability P is the analogue of the Born-layer observable. The requirement that P be resolution-independent is the analogue of the rendering-stability condition.

The CNN analogy also makes clear why a “renormalisable” theory is special: it is one where only finitely many operators (the relevant and marginal ones) need to be tuned, because the irrelevant ones automatically wash out. The ϕ^4 theory in $d = 4$ has three relevant/marginal parameters (m^2 , λ , Z); all others are irrelevant. A “non-renormalisable” theory (like gravity) has infinitely many marginal or relevant operators, meaning every resolution change introduces new parameters that must be measured—but it is still predictive at low energies, because the irrelevant operators are suppressed.

22.2 The Euclidean generating functional on the lattice

We work in Euclidean signature throughout this section: Wick-rotated, positive weight, well-defined measure, absolutely convergent. This is the natural setting for renormalization because the integrals converge without $i\varepsilon$ prescriptions, and Monte Carlo sampling is directly applicable for non-perturbative verification.

The Euclidean generating functional on the lattice. On a d -dimensional hypercubic lattice with N sites and spacing a :

$$\boxed{Z_E[J; \theta(a), a] = \int_{\mathbb{R}^N} \prod_n d\phi_n e^{-S_E[\phi; \theta(a), a] + a^d \sum_n J_n \phi_n}.} \quad (22.2)$$

The Euclidean lattice action for ϕ^4 theory:

$$S_E[\phi; \theta(a), a] = a^d \sum_n \left[\frac{Z(a)}{2} \sum_{\mu=1}^d \left(\frac{\phi_{n+\hat{\mu}} - \phi_n}{a} \right)^2 + \frac{m^2(a)}{2} \phi_n^2 + \frac{\lambda(a)}{4!} \phi_n^4 \right]. \quad (22.3)$$

Every quantity in this expression is finite:

The number of integration variables:	$N = N_\tau N_s^3$ (finite)
Each integration variable:	$\phi_n \in \mathbb{R}$
The action S_E :	A polynomial in ϕ_n (finite for finite ϕ)
The weight e^{-S_E} :	Positive, bounded above by 1 (at $\phi = 0$)
The integral:	Absolutely convergent (for $\lambda(a) > 0$)

The lattice is the engine. The lattice spacing a parametrises the engine's resolution. The engine parameters $\theta(a) = (m^2(a), \lambda(a), Z(a))$ are the coefficients of the action at that resolution.

Step 1: compute $\mathcal{O}[\theta(a), a]$. All Fock-layer quantities are expectations in the positive probability distribution $P[\phi] \propto e^{-S_E[\phi; \theta(a), a]}$:

$$G_E^{(n)}(x_1, \dots, x_n; \theta(a), a) = \frac{\int \prod_m d\phi_m \phi(x_1) \cdots \phi(x_n) e^{-S_E}}{\int \prod_m d\phi_m e^{-S_E}}. \quad (22.4)$$

These are ordinary statistical-mechanics expectation values, computable by perturbation theory (Feynman diagrams) or by Monte Carlo sampling. Each correlator is a function of $\theta(a)$ and a .

Born-layer observables are extracted from these correlators via the bridges:

$$m_{\text{phys}}^2[\theta(a), a] = \text{zero of } \Gamma^{(2)}(p; \theta(a), a), \quad (22.5)$$

$$\lambda_{\text{phys}}[\theta(a), a] = \Gamma^{(4)}(s_0, t_0, u_0; \theta(a), a), \quad (22.6)$$

etc. At this stage, the Born observables depend on both $\theta(a)$ and a .

Step 2: impose rendering stability. Demand that the Born outputs are a -independent:

$$m_{\text{phys}}^2[\theta(a), a] = m_{\text{phys}}^2 + o(a/\lambda), \quad (22.7)$$

$$\lambda_{\text{phys}}[\theta(a), a] = \lambda_{\text{phys}} + o(a/\lambda), \quad (22.8)$$

$$Z_{\text{phys}}[\theta(a), a] = 1 + o(a/\lambda). \quad (22.9)$$

These three equations determine $m^2(a)$, $\lambda(a)$, and $Z(a)$ as functions of a and the three physical observables $(m_{\text{phys}}, \lambda_{\text{phys}}, 1)$. The engine parameters absorb all the a -dependence so the Born outputs are stable.

Step 3: verify renormalisability. Having fixed three engine parameters by three conditions, check that *every* Born-layer observable is automatically a -independent. For ϕ^4 in $d = 4$: the a -dependent structures in $\Gamma^{(n)}$ for $n \geq 6$ are suppressed by powers of $(a/\lambda)^{n-4}$ (they correspond to irrelevant operators of dimension > 4). Three engine parameters suffice.

For gravity (a non-renormalisable theory): each loop order introduces new a -dependent structures requiring new engine parameters. The theory is still perfectly well-defined at finite a ; it is non-renormalisable only in the sense that infinitely many parameters are needed to maintain rendering stability as $a \rightarrow 0$.

The rendering-stability condition, restated in path-integral language. The connected generating functional $W_E[J; \theta(a), a] = \ln Z_E[J; \theta(a), a]$ must satisfy:

$$\boxed{W_E[J; \theta(a), a] = W_E^{\text{phys}}[J] + o(a/\lambda),} \quad (22.10)$$

for all slowly varying sources J (wavelengths $\lambda \gg a$). Since W_E generates all connected correlators, this single condition implies the stability of every $G_{\text{conn}}^{(n)}$, every $\Gamma^{(n)}$, and hence every Born-layer observable.

22.3 The one-loop effective potential: a worked example

The effective potential $V_{\text{eff}}(\bar{\phi})$ is Γ_E evaluated at a constant field $\bar{\phi}$: $\Gamma_E[\bar{\phi} = \text{const}] = \Omega \cdot V_{\text{eff}}(\bar{\phi})$, where $\Omega = L^d$ is the spacetime volume. It determines the vacuum structure of the theory: the physical vacuum is the global minimum of V_{eff} .

The tree-level potential. At tree level, $\Gamma \approx S$, so:

$$V_{\text{tree}}(\bar{\phi}) = \frac{m^2(a)}{2} \bar{\phi}^2 + \frac{\lambda(a)}{4!} \bar{\phi}^4. \quad (22.11)$$

If $m^2(a) > 0$: a single minimum at $\bar{\phi} = 0$ (unbroken symmetry). If $m^2(a) < 0$: two minima at $\bar{\phi} = \pm v$ with $v^2 = -6m^2(a)/\lambda(a)$ (spontaneous symmetry breaking).

The one-loop correction. The one-loop effective action (21.27) gives:

$$V_{\text{eff}}(\bar{\phi}) = V_{\text{tree}}(\bar{\phi}) + V^{(1)}(\bar{\phi}) + O(\lambda^2), \quad (22.12)$$

where the one-loop correction is the integral:

$$V^{(1)}(\bar{\phi}) = \frac{1}{2} \int_{|k| < \pi/a} \frac{d^d k}{(2\pi)^d} \ln[k^2 + M^2(\bar{\phi}; a)], \quad M^2(\bar{\phi}; a) = m^2(a) + \frac{\lambda(a)}{2} \bar{\phi}^2. \quad (22.13)$$

The momentum integral runs over the lattice Brillouin zone $|k| < \pi/a$: the lattice spacing provides the UV cutoff. The integrand is the logarithm of the fluctuation operator's eigenvalue at momentum k : each mode contributes $\frac{1}{2} \ln(k^2 + M^2)$ to the vacuum energy (the “zero-point energy” of an oscillator with frequency $\omega_k = \sqrt{k^2 + M^2}$).

Explicit evaluation in $d = 4$. Introduce $\Lambda = \pi/a$. The integral can be evaluated by standard techniques (dimensional regularisation gives the same result in the $a \rightarrow 0$ limit, but we stay on the lattice):

$$\begin{aligned} V^{(1)}(\bar{\phi}) &= \frac{1}{2} \int_0^\Lambda \frac{4\pi k^3 dk}{(2\pi)^4} \ln(k^2 + M^2) \\ &= \frac{1}{64\pi^2} \left[\Lambda^4 \ln\left(1 + \frac{M^2}{\Lambda^2}\right) - M^2 \Lambda^2 + \frac{M^4}{2} \ln \frac{\Lambda^2}{M^2} + \dots \right], \end{aligned} \quad (22.14)$$

where the ellipsis denotes terms that are finite as $\Lambda \rightarrow \infty$ (and hence $a \rightarrow 0$). The potentially problematic terms are:

The Λ^4 term: a quartic dependence on the cutoff. This contributes a constant to V_{eff} , independent of $\bar{\phi}$, and can be absorbed into the cosmological constant. It does not affect the location of the minimum of V_{eff} and has no Born-layer consequence for the scalar field theory (though it would matter for gravity).

The $M^2 \Lambda^2 = [m^2(a) + \frac{\lambda(a)}{2} \bar{\phi}^2] \Lambda^2$ term: a quadratic dependence on the cutoff that is $\bar{\phi}$ -dependent. It shifts the mass: $\delta m^2 \sim \lambda \Lambda^2 / 16\pi^2 = \lambda \pi^2 / (16\pi^2 a^2)$. This must be absorbed into $m^2(a)$.

The $M^4 \ln(\Lambda^2/M^2)$ term: a logarithmic dependence on the cutoff. Expanding $M^4 = [m^2 + \frac{\lambda}{2} \bar{\phi}^2]^2$ produces terms proportional to $\bar{\phi}^4$, which shift the coupling: $\delta \lambda \sim \lambda^2 \ln \Lambda / 16\pi^2$. This must be absorbed into $\lambda(a)$.

The stability condition in action. The physical effective potential is $V_{\text{eff}}^{\text{phys}}(\bar{\phi}) = V_{\text{eff}}(\bar{\phi}; a, \theta(a))$ where $\theta(a)$ is chosen so that the result is a -independent (to accuracy $o(a/\lambda)$). Concretely, we demand:

$$V_{\text{eff}}^{\text{phys}''}(0) = m_{\text{phys}}^2 \quad (\text{fixes } m^2(a): \text{ absorbs the } \Lambda^2 \text{ term}), \quad (22.15)$$

$$V_{\text{eff}}^{\text{phys}''''}(0) = \lambda_{\text{phys}} \quad (\text{fixes } \lambda(a): \text{ absorbs the } \ln \Lambda \text{ term}). \quad (22.16)$$

These two conditions determine $m^2(a)$ and $\lambda(a)$ as functions of a , in terms of the physical mass m_{phys} and the physical coupling λ_{phys} . The physical effective potential $V_{\text{eff}}^{\text{phys}}$ is then a -independent: two cinemas, same movie.

The Coleman–Weinberg potential. A famous application: set $m^2(a) = 0$ (the classically massless case). At tree level, the potential is flat near $\bar{\phi} = 0$: no symmetry breaking. But the one-loop correction introduces a logarithmic term:

$$V_{\text{CW}}(\bar{\phi}) = \frac{\lambda}{4!} \bar{\phi}^4 + \frac{\lambda^2}{256\pi^2} \bar{\phi}^4 \left(\ln \frac{\bar{\phi}^2}{v^2} - \frac{3}{2} \right), \quad (22.17)$$

where v is a renormalization scale. The logarithmic correction creates a new minimum at $\bar{\phi} \neq 0$: quantum fluctuations (the one-loop correction) break the symmetry that the classical potential preserves. This is *radiative symmetry breaking*: the Born-layer vacuum (the minimum of the physical V_{eff}) differs from the tree-level vacuum because the loop correction tips the potential.

The Coleman–Weinberg potential is a Born-layer prediction: it determines the vacuum expectation value v (which sets particle masses through the Higgs mechanism), and the mass of the scalar excitation around the minimum. Both are measurable. The calculation is Fock-layer (a one-loop integral on the lattice); the result is Born-layer (a VEV, a mass).

22.4 Renormalization conditions on $\Gamma^{(n)}$

The rendering-stability condition (??) on W_E translates to specific constraints on the 1PI vertex functions $\Gamma^{(n)}$ generated by the effective action $\Gamma[\bar{\phi}]$.

The logic is the same as on the Hamiltonian side (section 20.7), but we now have the explicit one-loop expressions from the path integral to work with.

The inverse propagator. The two-point vertex function in momentum space:

$$\Gamma^{(2)}(p; a) = Z(a)p^2 - m^2(a) - \Sigma(p; a), \quad (22.18)$$

where $\Sigma(p; a)$ is the self-energy, computed from Feynman diagrams with the lattice cutoff $|k| < \pi/a$. The stability condition requires three things:

Condition 1: the pole position is a -independent.

The physical mass is the zero of $\Gamma^{(2)}(p)$:

$$\Gamma_R^{(2)}(p^2 = m_{\text{phys}}^2) = 0. \quad (22.19)$$

At one loop, using the tadpole self-energy (21.33):

$$m_{\text{phys}}^2 = m^2(a) + \Sigma^{(1)}(a) = m^2(a) + \frac{\lambda(a)}{32\pi^2} \left[\frac{\pi^2}{a^2} - m^2(a) \ln \frac{\pi/a}{m(a)} + \dots \right]. \quad (22.20)$$

This determines $m^2(a)$ as a function of a :

$$m^2(a) = m_{\text{phys}}^2 - \frac{\lambda(a)}{32\pi^2} \cdot \frac{\pi^2}{a^2} + O(\lambda m^2 \ln(1/a)). \quad (22.21)$$

As $a \rightarrow 0$: $m^2(a) \rightarrow -\infty$. The bare mass is large and negative at small lattice spacing. This is not a sickness; it is the engine’s internal calibration. The Born-layer mass m_{phys} is the one that stays fixed.

Condition 2: the residue is normalised.

The derivative of $\Gamma_R^{(2)}$ at the mass shell:

$$\left. \frac{d\Gamma_R^{(2)}}{dp^2} \right|_{p^2=m_{\text{phys}}^2} = 1. \quad (22.22)$$

This determines $Z(a)$: the wave-function renormalization. The renormalised field $\hat{\phi}_R = \hat{\phi}/\sqrt{Z(a)}$ has a propagator with residue 1 at the physical mass—independent of a .

Condition 3: the coupling is a -independent.

The four-point vertex at a symmetric momentum configuration (the “symmetric point” with $p_i \cdot p_j = (4m^2 - s)/3$ for $i \neq j$, $s = 4m^2$):

$$\Gamma_R^{(4)}|_{\text{sym. pt}} = \lambda_{\text{phys}}. \quad (22.23)$$

This determines $\lambda(a)$ as a function of a .

Summary. Three Born-layer numbers— m_{phys} , $Z = 1$ (normalisation convention), λ_{phys} —fix three bare parameters $m^2(a)$, $Z(a)$, $\lambda(a)$ at each lattice spacing a . The theory is renormalisable because three conditions suffice for *all* n -point functions: once $\Gamma^{(2)}$ and $\Gamma^{(4)}$ are stable, every $\Gamma^{(n)}$ (and hence every Born-layer observable) is automatically stable. This is a non-trivial result: it requires that all a -dependent terms in $\Gamma^{(n)}$ for $n > 4$ are already absorbed by the running of m^2 , λ , and Z .

22.5 The beta function from the path integral

The beta function is the rate at which the coupling runs as the lattice spacing changes:

$$\beta(\lambda) = \frac{d\lambda}{d \ln(1/a)} = a \frac{d\lambda}{da} \Big|_{a < 0}. \quad (22.24)$$

On the path-integral side, the beta function can be extracted from the one-loop vertex correction (21.34). The a -dependent part of $\Gamma^{(4)}$ at one loop:

$$\delta\Gamma^{(4)}|_{a\text{-dep}} = \frac{3\lambda^2}{16\pi^2} \ln \frac{\pi/a}{\mu}, \quad (22.25)$$

where the factor of 3 comes from the three channels (s, t, u) and μ is the renormalization scale. The stability condition (22.23) absorbs this into $\lambda(a)$: $\lambda(a) = \lambda_{\text{phys}} + \frac{3\lambda_{\text{phys}}^2}{16\pi^2} \ln(\pi/(a\mu)) + O(\lambda^3)$. Differentiating:

$$\boxed{\beta(\lambda) = \frac{3\lambda^2}{16\pi^2} + O(\lambda^3)}. \quad (22.26)$$

This agrees with:

The Hamiltonian-side beta function (20.28), extracted from the Feshbach coupling shift (??).

The heat-kernel beta function (23.27), extracted from the Seeley–DeWitt coefficient a_2 .

Three computational methods—operator-side Feshbach, path-integral vertex correction, heat-kernel coefficient—give the same beta function. This is a consistency check on the rendering-stability framework: the Born-layer prediction (how the coupling runs) does not depend on which computational language the engine speaks.

What the beta function means. $\beta(\lambda) > 0$ means $\lambda(a)$ increases as a decreases (finer resolution): the engine parameter must be made stronger at finer lattice spacing to maintain the same Born-layer outputs. This is a statement about the engine’s internal calibration, not a Born-layer statement. No observer measures $\lambda(a)$ directly—the observer measures a cross section σ at a specific energy, and $\lambda_{\text{phys}}(\mu)$ is defined by fitting σ to the theoretical prediction.

The *running coupling* $\lambda_{\text{phys}}(\mu)$ as a function of the observation energy μ is governed by the same beta function, and this running IS Born-layer: it describes how the measured cross section changes with experimental energy. At low μ : the coupling is small, perturbation theory works. At high μ (approaching the Landau pole): the coupling grows, and perturbation theory breaks down.

In the lattice-first view, the Landau pole is not a catastrophe: the lattice spacing a provides a resolution floor, the engine parameter $\lambda(a)$ is a finite number at that floor, and the “pole” is an artefact of extrapolating the perturbative beta function beyond its domain of validity.

The Hamiltonian–path-integral agreement. The beta function $\beta(\lambda) = 3\lambda^2/(16\pi^2) + O(\lambda^3)$ was obtained three ways:

Method	Section	Computation
Hamiltonian (Feshbach)	§20.5	Shell integral of $H_{PQ}(1/H_{QQ})H_{QP}$
Path integral (vertex)	§22.5	One-loop $\delta\Gamma^{(4)}$ on the lattice
Heat kernel	§23.12	Seeley–DeWitt coefficient a_2

All three give the same number. This is the consistency check: the Born-layer prediction (how the coupling runs) does not depend on which Fock-layer computational language the engine speaks. The engine parameters $\theta(a)$ are the same on both sides; the Born-layer observables are the same; the rendering-stability condition is the same. Only the method of computation differs.

22.6 The correspondence: Hamiltonian and path integral

The two treatments of renormalization are two views of the same rendering-stability condition. The following table records the dictionary:

Hamiltonian side	Path integral side
$H(a)$ on $\mathcal{F}_a$	$S_E[\phi; a, \theta(a)]$ on lattice
Feshbach effective H_{eff} on $\mathcal{F}_{a'}$	Wilsonian effective action $S_{\text{eff}}[\phi_<]$ after integrating out $\phi_>$
Wegner flow $dH/ds = [[\eta, H], H]$	Exact RG: Polchinski $\partial_\Lambda S_\Lambda = \frac{1}{2} \text{Tr} [\dot{G}(\delta^2 S / \delta \phi^2 - (\delta S / \delta \phi)^2)]$
Feshbach coupling shift $\delta\lambda = -3\lambda^2 \int_{\text{shell}} \dots$	One-loop vertex correction $\delta\Gamma^{(4)} = 3\lambda^2 \int \dots$
$\beta_H(\lambda) = d\lambda/d\ln(1/a)$	$\beta_{\text{PI}}(\lambda) = d\lambda/d\ln\mu$
Spectral gap $= m_{\text{phys}}$	Pole of $\Gamma^{(2)}(p)^{-1}$
$Z(a) = \langle \Omega \hat{\phi} \mathbf{p} \rangle ^2$	Residue of propagator pole
Heat kernel a_2 coefficient	One-loop $\int d^4k \ln(k^2 + M^2)$
Same $\theta(a)$, same $\beta(\lambda)$	Same Born-layer observables

Row 3 deserves comment. The Wegner flow on the Hamiltonian side and the Polchinski equation on the path-integral side are both *exact* RG flows: they do not rely on perturbation theory. The Wegner flow is a unitary transformation that band-diagonalises H continuously; the Polchinski equation evolves the effective action S_Λ as modes are integrated out. Both preserve all eigenvalues (the spectrum of H , equivalently the poles of $\Gamma^{(2)}$) and all Born-layer predictions. They differ in formulation (operators vs. functionals) but agree on everything physical.

22.7 Why both sides exist

If the two sides give the same answers, why maintain both? Because each has computational strengths that the other lacks.

The Hamiltonian side is conceptually primary. The engine is the Hamiltonian $H(a)$, the Schrödinger equation $\dot{h}_t = -iHh_t$ is the update rule, and the Fock space $\mathcal{F}_a$ is the state space. The Hamiltonian side works for any theory (no sign problem: the operator H is well-defined regardless of whether $e^{-S_E} \geq 0$), connects directly to the Fock-Born architecture, and makes the operator-level structure (correlators, spectral functions, the heat kernel) transparent. The Feshbach–Löwdin projection is an exact, non-perturbative computation on the operator side; translating it to the path integral produces the Wilsonian effective action, which is less transparent.

The path-integral side is computationally efficient for gauge theories. The Faddeev–Popov procedure for fixing gauge redundancy (section ??) is natural in path-integral language: one restricts the domain of integration by inserting a gauge-fixing delta function and a Faddeev–Popov determinant. Translating this to the Hamiltonian side requires the Gupta–Bleuler or BRST formalism (section ??), which is more involved.

The path integral makes Lorentz invariance manifest: the Euclidean action S_E is $SO(4)$ -invariant, while the Hamiltonian formulation breaks this by singling out time. For computations where Lorentz invariance simplifies the algebra (dimensional regularisation, the background-field method), the path integral is more efficient.

The Euclidean path integral enables Monte Carlo sampling: when $e^{-S_E} \geq 0$ (no sign problem), the factor e^{-S_E}/Z_E is a normalised probability distribution, and the generating functional can be

estimated by importance sampling. This is lattice QCD: the dominant source of non-perturbative predictions for hadron masses, the strong coupling α_s , and the QCD phase diagram.

When to use which. The choice is a choice of computational strategy, not of physics. Both are Fock-layer tools. Both serve the same Born-layer predictions.

Use the Hamiltonian when...	Use the path integral when...
The operator structure is needed	Lorentz invariance must be manifest
The theory has a sign problem	Monte Carlo sampling is available
The spectrum is the target	Gauge-fixing is needed
The Fock-Born architecture is the focus	Symmetry arguments dominate
Non-perturbative operator identities help	Perturbative diagrams are efficient

23 Wick Rotation, Quantum Monte Carlo, and the Heat Kernel

The real-time path integral $\int \mathcal{D}\phi e^{iS_M}$ has an oscillatory integrand of unit modulus. The Wick rotation $t \rightarrow -i\tau$ replaces e^{iS_M} with e^{-S_E} : an absolutely convergent integral with positive weights, sampleable by Monte Carlo. This is a change of computational strategy inside the Fock layer, not a change of physics.

The Euclidean path integral also gives access to the *heat kernel*—the Euclidean propagator $e^{-\tau D}$, given its proper name and developed as a systematic tool. The heat kernel, introduced by Schwinger in his proper-time method (1951), provides a universal operator-side technology for extracting masses, UV divergences, beta functions, and anomalies without Feynman diagrams.

23.1 Wick rotation for the scalar field

Under $t = -i\tau$: $dt = -id\tau$, $\dot{\phi}^2 \rightarrow -(\partial_\tau \phi)^2$. The Euclidean action:

$$S_E = \int d^4x_E \left[\frac{1}{2}(\partial_\mu^E \phi)^2 + \frac{1}{2}m^2 \phi^2 + \frac{\lambda}{4!} \phi^4 \right], \quad (23.1)$$

with $(\partial_\mu^E \phi)^2 = (\partial_\tau \phi)^2 + (\nabla \phi)^2$. All terms positive for $\lambda > 0$: absolutely convergent. The Minkowski $SO(3, 1)$ becomes the Euclidean $SO(4)$.

23.2 Euclidean correlators and their Lorentzian counterparts

The Euclidean generating functional $Z_E[J] = \int \mathcal{D}\phi e^{-S_E + \int J\phi}$ and the connected generating functional $W_E[J] = \ln Z_E[J]$ generate Euclidean correlators: $G_E^{(n)} = \delta^n W_E / \delta J^n|_{J=0}$. These are the same correlators as on the Hamiltonian side (section 18.3), analytically continued to imaginary time.

The spectral representation in Euclidean time:

$$G_E^{(2)}(\tau, \mathbf{p}) = \int_0^\infty d\mu \rho(\mu) e^{-\mu|\tau|}, \quad (23.2)$$

where $\rho(\mu)$ is the same spectral function as in the Lorentzian Källén–Lehmann representation (18.38). Euclidean: Laplace transform of ρ . Lorentzian: Fourier transform of ρ . Both encode the same spectrum.

At large τ : $G_E^{(2)}(\tau, \mathbf{0}) \sim Z e^{-m_{\text{phys}}\tau}$. The mass is the exponential decay rate of the Euclidean correlator.

23.3 The Euclidean shadow: a catalog of correspondences

The Euclidean theory is not a separate theory. It is a *twisted shadow* of the Lorentzian theory: a projection obtained by rotating the time contour from the real axis to the imaginary axis. The shadow preserves some features of the original (the spectrum, the symmetries, the operator content) and distorts others (the time evolution, the causal structure, the real-time dynamics). It is almost a miracle that one can reconstruct the original from the shadow—and in fact, one cannot always do so.

The analytic continuation. The passage from Lorentzian to Euclidean is the substitution $k^0 \rightarrow ik_E^0$ (equivalently $t \rightarrow -i\tau$). This maps the Lorentzian propagator to the Euclidean one:

$$\frac{i}{(k^0)^2 - |\mathbf{k}|^2 - m^2 + i\varepsilon} \xrightarrow{k^0 = ik_E^0} \frac{-1}{k_E^{02} + |\mathbf{k}|^2 + m^2} = \frac{-1}{k_E^2 + m^2}. \quad (23.3)$$

The Lorentzian propagator has poles at $k^0 = \pm\sqrt{|\mathbf{k}|^2 + m^2}$ (on-shell particles); the Euclidean propagator has no poles (it is strictly negative for all real k_E). The poles have moved off the real axis into the complex plane. The spectrum is encoded, but hidden.

The inverse passage—from Euclidean back to Lorentzian—requires analytic continuation $k_E^0 \rightarrow -ik^0$, which must navigate around the poles. This is the Osterwalder–Schrader reconstruction: under certain axioms (reflection positivity, cluster decomposition), the Euclidean correlators uniquely determine the Lorentzian theory.

The catalog. The following table lists what can and cannot be extracted from the Euclidean theory, with the method and the difficulty.

Born-layer observable	Euclidean route	Method	Difficulty	Status
Masses m_n	$G_E(\tau) \sim e^{-m_n\tau}$	Exponential fit at large τ	Easy	Routine
Energy gaps ΔE	$G_E(\tau) \propto e^{-\Delta E\tau}$	Correlator ratio	Easy	Routine
Ground-state wavefn.	Histogram of ϕ at large β	Monte Carlo sampling	Easy	Routine
Vacuum condensates	$\langle\bar{\psi}\psi\rangle_E, \langle F^2\rangle_E$	Direct measurement	Easy	Routine
Scattering lengths	Finite-volume energy shifts $\delta E(L)$	Lüscher formula	Medium	Established
Phase shifts $\delta(k)$	Multi-level Lüscher analysis	Generalised eigenvalue problem	Medium	Active research
Form factors	3-point functions at large τ	Ratio method	Medium	Established
Thermodynamics	$\mathcal{Z}(\beta)$ with periodic BC	$F = -\ln \mathcal{Z}/\beta$	Easy	Routine
Beta functions	Step-scaling; Seeley–DeWitt a_2	Heat kernel or gradient flow	Easy	Routine
Spectral function $\rho(\omega)$	Inverse Laplace of $G_E(\tau)$	MEM, Backus–Gilbert	Hard	Ill-posed
Transport coefficients	Kubo formula requires $\rho(\omega \rightarrow 0)$	Analytic continuation	Very hard	Unreliable
Real-time dynamics	Not available	N/A	–	Inaccessible
Out-of-equilibrium	Not available	N/A	–	Inaccessible
Finite-density QCD	$\det M$ complex at $\mu_B \neq 0$	Reweighting, Taylor expansion	Very hard	Sign problem

The miracle. The easy rows are remarkable. The Euclidean theory has no time evolution, no causality, no particles, no scattering—it is a classical statistical system on a d -dimensional lattice. Yet from this static, Euclidean object we can extract the Lorentzian mass spectrum, the scattering amplitudes, the anomalies, and the beta functions of the original quantum theory. The spectrum is hidden in the exponential decay rates of Euclidean correlators. The scattering amplitudes are hidden in the finite-volume energy shifts. The anomalies are hidden in the short-distance expansion of the heat kernel. All of this information is encoded in the Euclidean

shadow, recoverable by analytic continuation—provided the analytic continuation is well-defined.

The limitation. The hard and inaccessible rows are equally important. The Euclidean theory does not know about real-time dynamics: there is no time-evolution operator e^{-iHt} , no causal ordering, no retardation. Information that lives specifically in the Lorentzian time structure—transport coefficients, thermalisation rates, out-of-equilibrium phenomena—is either exponentially hard to extract (the inverse Laplace problem) or flatly inaccessible (no Euclidean formulation of real-time dynamics exists).

This is where the Hamiltonian comes back. The Hamiltonian $H(a)$ —a Hermitian matrix on $\mathcal{F}_a$ —operates in Lorentzian time by definition. It does not need Wick rotation. It does not have a sign problem. It computes real-time dynamics, transport coefficients, and out-of-equilibrium evolution directly. The limitation is computational: exact diagonalisation scales exponentially, and tensor-network methods are limited to low dimensions or special structures. But the limitation is in our computational power, not in the formalism. The Hamiltonian has access to everything; the Euclidean path integral has access to a (remarkably informative) subset.

In ML terms. The Euclidean theory is like a set of sufficient statistics for a family of probability distributions: it compresses the original (Lorentzian) theory into a lower-dimensional representation that preserves some information (the spectrum, the scattering parameters) and discards other information (the real-time dynamics, the causal structure). The compression is lossy: you can reconstruct the masses but not the real-time correlators. The heat kernel is the tool that reads the compressed representation.

23.4 The lattice Euclidean path integral

On a lattice with spacing a and $N_\tau \times N_s^3$ sites:

$$\mathcal{Z}(a) = \int \prod_n d\phi_n e^{-S_E[\{\phi_n\};a]}. \quad (23.4)$$

This is a finite-dimensional integral with positive weight $e^{-S_E} \geq 0$: a classical statistical mechanics partition function.

Euclidean QFT on lattice	Classical statistical mechanics
Field ϕ_n at site n	Spin/displacement at site n
Euclidean action S_E	Energy $E/k_B T$
e^{-S_E}	Boltzmann weight $e^{-E/k_B T}$
Euclidean time extent $N_\tau a$	Inverse temperature β
Lattice spacing a	UV cutoff
Correlation length ξ	$\xi = 1/m_{\text{phys}}$
Continuum limit $a \rightarrow 0$	Critical point $\xi/a \rightarrow \infty$

The continuum limit of the lattice field theory is the critical point of the classical statistical system: the lattice spacing a becomes invisible when the correlation length $\xi = 1/m_{\text{phys}} \gg a$. This is the connection between renormalization and critical phenomena that Wilson identified.

23.5 Quantum Monte Carlo: what it computes

The positive weight $e^{-S_E} \geq 0$ makes Monte Carlo sampling possible: generate field configurations $\{\phi_n^{(s)}\}$ ($s = 1, \dots, N_{\text{samples}}$) from the distribution $P[\{\phi_n\}] \propto e^{-S_E}$ using Metropolis or hybrid Monte Carlo, and estimate expectation values as sample averages.

Masses from correlator decay. Measure the Euclidean two-point function:

$$C_E(\tau) = \frac{1}{N_{\text{samples}}} \sum_s \phi^{(s)}(\tau, \mathbf{0}) \phi^{(s)}(0, \mathbf{0}) \xrightarrow{\tau \rightarrow \infty} Z e^{-m_{\text{phys}}\tau}. \quad (23.5)$$

Fit the exponential decay to extract m_{phys} . This is how lattice QCD computes the proton mass, the pion mass, and the entire hadron spectrum: each hadron is a distinct exponential in $C_E(\tau)$, with its own decay rate. Higher excited states decay faster; the ground state dominates at large τ .

Matrix elements from three-point functions. $\langle \phi(\tau_1) \mathcal{O}(\tau_2) \phi(\tau_3) \rangle_E$ at large $|\tau_1 - \tau_2|$ and $|\tau_2 - \tau_3|$ gives the matrix element $\langle n | \mathcal{O} | m \rangle$ between states n and m . This is how form factors and weak-decay matrix elements are computed.

Scattering via the Lüscher method. On a finite spatial lattice of size L , the two-particle energy levels shift from the non-interacting values by an amount that depends on the scattering phase shift $\delta(k)$. Lüscher’s formula (1986):

$$\delta E_n(L) = -\frac{4\pi a_s}{m_{\text{phys}} L^3} \left(1 + c_1 \frac{a_s}{L} + c_2 \frac{a_s^2}{L^2} + \cdots \right), \quad (23.6)$$

where a_s is the scattering length and c_1, c_2 are known geometric constants. The procedure: measure two-particle energy levels at several values of L by Monte Carlo, extract $\delta E_n(L)$, fit to the Lüscher formula, and read off the scattering parameters. This converts Euclidean energy levels (Fock-layer, computable by Monte Carlo) into scattering parameters (Born-layer, measurable in experiment).

The spectral function: an inverse problem. Reconstructing the full spectral function $\rho(\mu)$ from $G_E^{(2)}(\tau) = \int d\mu \rho(\mu) e^{-\mu\tau}$ is an ill-posed inverse Laplace transform. The exponential kernel $e^{-\mu\tau}$ smears out fine structure in ρ : many different spectral functions produce similar Euclidean correlators. Methods (maximum entropy, Backus–Gilbert, stochastic reconstruction) provide controlled approximations, but the difficulty is intrinsic. The Euclidean signature gives easy access to masses (exponential decay rates) but hard access to spectral details (inverse problems). This is a Fock-layer computational limitation, not a Born-layer physics limitation.

Thermodynamics. The partition function $\mathcal{Z}(\beta) = \text{Tr } e^{-\beta H}$ gives free energy $F = -\ln \mathcal{Z}/\beta$, pressure, energy density, and susceptibilities by differentiation. These are Born-layer thermodynamic observables computed from the Euclidean path integral.

23.6 The sign problem

The positive weight $e^{-S_E} \geq 0$ fails in several important cases:

Fermions: Grassmann integration produces a determinant $\det M$ that can be negative. In lattice QCD at zero chemical potential, $\det M$ is positive (after doubling tricks), and Monte Carlo works. At finite baryon density $\mu_B \neq 0$, the determinant becomes complex and Monte Carlo breaks down.

Real-time dynamics: the real-time weight e^{iS_M} is oscillatory by definition; no Wick rotation can help when the physical question involves real-time evolution (transport coefficients, thermalisation, out-of-equilibrium dynamics).

The sign problem is a Fock-layer computational obstacle, not a physical one. The Hamiltonian is perfectly well-defined regardless of whether the Euclidean path integral can be sampled. This is where Hamiltonian methods—exact diagonalisation, DMRG, tensor networks, quantum simulation—become essential. They work directly with H and never construct a probability measure to sample from.

23.7 What the Euclidean side cannot see

The Euclidean shadow is powerful but incomplete. There are Born-layer observables that can be computed directly on the Lorentzian (Hamiltonian) side but are exponentially hard or impossible to extract from the Euclidean path integral. Honesty requires cataloguing them.

1. Real-time dynamics. The Euclidean path integral has no time-evolution operator. The weight e^{-S_E} generates a thermal equilibrium distribution, not a dynamical evolution. Questions about how a system evolves in real time—how a quark-gluon plasma thermalises, how a heavy ion collision unfolds, how a quantum quench relaxes—have no direct Euclidean formulation. The Hamiltonian e^{-iHt} answers these questions; the Euclidean path integral does not.

2. Transport coefficients. Shear viscosity η , electrical conductivity σ_e , and other transport coefficients are defined by the Kubo formula: $\eta = \lim_{\omega \rightarrow 0} \rho_{T_{12}T_{12}}(\omega)/(2\omega)$, where $\rho(\omega)$ is the spectral function of the stress tensor. Extracting $\rho(\omega)$ from the Euclidean correlator $G_E(\tau) = \int d\omega \rho(\omega) K(\tau, \omega)$ requires an inverse Laplace transform—an ill-posed problem. The exponential kernel $K(\tau, \omega) = \cosh[\omega(\beta/2 - \tau)]/\sinh(\omega\beta/2)$ smears out the low- ω structure of ρ that determines transport coefficients. In practice, the extraction is unreliable: many different $\rho(\omega)$ fit the same $G_E(\tau)$ within Monte Carlo noise.

3. The spectral function at finite resolution. Even the full spectral function $\rho(\omega)$ (not just the low- ω limit) is hard to extract. The inverse Laplace transform is exponentially ill-conditioned: the singular values of the kernel $K(\tau, \omega)$ decay exponentially, so small noise in $G_E(\tau)$ produces large errors in $\rho(\omega)$. Methods (Maximum Entropy, Backus–Gilbert, stochastic analytical continuation) provide controlled approximations, but the resolution in ω is fundamentally limited by the statistical precision of the Euclidean data.

By contrast, on the Hamiltonian side: the spectral function is directly accessible as $\rho(\omega) = \sum_n |\langle n | \hat{O} | 0 \rangle|^2 \delta(\omega - E_n)$ —a list of delta functions at the eigenvalues of H , with weights given by matrix elements. No inverse problem; no ill-conditioning; exact in principle.

4. Finite-density QCD. At finite baryon chemical potential μ_B , the fermion determinant $\det M(\mu_B)$ becomes complex: $e^{-S_E} \det M$ is no longer positive, and Monte Carlo sampling breaks down (the sign problem). The QCD phase diagram at $\mu_B \neq 0$ —the location of the critical point, the nature of the deconfinement transition—is inaccessible to standard lattice Monte Carlo.

The Hamiltonian side does not have a sign problem: $H(\mu_B)$ is a well-defined Hermitian operator (after a straightforward modification), and its eigenvalues and eigenstates can be computed by Hamiltonian methods (exact diagonalisation, DMRG, tensor networks, quantum simulation). The bottleneck is the exponential scaling of the Hilbert space, not the sign of a weight.

5. Minkowski-signature scattering. The Euclidean path integral does not produce scattering amplitudes directly. Scattering is a Lorentzian concept: it involves in-states at $t \rightarrow -\infty$, out-states at $t \rightarrow +\infty$, and the S -matrix $S = \lim_{T \rightarrow \infty} e^{iH_0 T} e^{-2iHT} e^{iH_0 T}$. The Euclidean path integral produces Euclidean correlators, which encode scattering information only after analytic continuation ($k_E^0 \rightarrow -ik^0$) and the LSZ procedure. On a finite lattice, the Lüscher method extracts scattering parameters from finite-volume energy shifts—a remarkable but indirect route.

The honest summary. The Euclidean path integral is the most powerful *non-perturbative computational tool* in the Fock-layer arsenal: it gives direct access to the spectrum, the vacuum structure, the thermodynamics, and the beta functions. But it is not the complete theory. The complete theory is the Hamiltonian $H(a)$ on the Fock space $\mathcal{F}_a$, evolving in Lorentzian time. The Euclidean path integral is a projection of this theory onto a tractable computational domain—a

projection that preserves the spectral information (the “eigenvalues”) but loses the dynamical information (the “time evolution”).

The heat kernel is the bridge: it reads the spectral information from the Euclidean shadow ($\tau \rightarrow \infty$ for masses, $\tau \rightarrow 0$ for UV structure) and converts it to Lorentzian predictions via analytic continuation. The limitations of the Euclidean side are the limitations of this bridge: what analytic continuation can recover, the Euclidean path integral provides; what it cannot, only the Hamiltonian can supply.

23.8 The heat kernel: the Euclidean propagator as a tool

The Euclidean time-evolution operator $e^{-\tau D}$, where $D = -\nabla^2 + V(x)$ (or more generally any elliptic operator on the field space), is the *heat kernel*. It solves the heat equation $\partial_\tau u = -Du$ in the same way that e^{-iHt} solves the Schrödinger equation. The Wick rotation $t \rightarrow -i\tau$ literally converts the Schrödinger propagator into the heat kernel.

The heat kernel is not a new object. It is the Euclidean propagator, given its proper name and developed as a systematic computational tool.

Spectral decomposition. If D has eigenvalues λ_n and eigenfunctions ψ_n :

$$K(x, y; \tau) \equiv \langle x | e^{-\tau D} | y \rangle = \sum_n e^{-\lambda_n \tau} \psi_n(x) \psi_n^*(y). \quad (23.7)$$

Two limits carry all the physics:

Large τ : the lowest eigenvalue dominates. $K(x, y; \tau) \sim e^{-\lambda_0 \tau} \psi_0(x) \psi_0^*(y)$. This is the ground-state projection—the same mechanism that extracts masses from Euclidean correlators (??).

Small τ : all eigenvalues contribute, and K probes the short-distance (UV) structure of D . This limit gives the Seeley–DeWitt expansion.

23.9 The Seeley–DeWitt expansion: UV structure without diagrams

As $\tau \rightarrow 0^+$, the heat kernel has the asymptotic expansion:

$$K(x, x; \tau) = \frac{1}{(4\pi\tau)^{d/2}} \sum_{n=0}^{\infty} a_n(x) \tau^n, \quad (23.8)$$

where d is the spacetime dimension and the *Seeley–DeWitt coefficients* $a_n(x)$ are local invariants built from the potential V , the curvature (if the manifold is curved), and the gauge field strength (if D is a gauge-covariant operator).

The first few coefficients for $D = -\nabla^2 + V(x)$ on flat space:

$$a_0(x) = 1, \quad (23.9)$$

$$a_1(x) = -V(x), \quad (23.10)$$

$$a_2(x) = \frac{1}{2}V(x)^2 + \frac{1}{6}\nabla^2 V(x). \quad (23.11)$$

For a gauge-covariant operator $D = -D_\mu D^\mu + E$ (where $D_\mu = \partial_\mu + iA_\mu$ is the covariant derivative and E is an endomorphism):

$$a_0 = \text{tr } \mathbf{1}, \quad (23.12)$$

$$a_1 = \text{tr } E, \quad (23.13)$$

$$a_2 = \text{tr} \left[\frac{1}{2}E^2 + \frac{1}{6}\nabla^2 E + \frac{1}{12}F_{\mu\nu}F^{\mu\nu} \right], \quad (23.14)$$

where $F_{\mu\nu} = [D_\mu, D_\nu]$ is the gauge field strength and tr is the trace over internal indices.

These coefficients encode the entire UV structure of the operator D . They are *local*: they depend only on V , $F_{\mu\nu}$, and their derivatives at the point x . They are *universal*: they depend on D but not on the boundary conditions, the topology, or the global structure of the manifold. And they are computable *without Feynman diagrams*: they follow from the short-time expansion of a single operator exponential.

23.10 Deriving the Seeley–DeWitt coefficients

The Seeley–DeWitt expansion (43.23) is not a conjecture; it is derived from the heat equation. We show the derivation for a general second-order operator $D = -g^{\mu\nu} D_\mu D_\nu + E$ on a d -dimensional Riemannian manifold, where $D_\mu = \nabla_\mu + iA_\mu$ is the gauge-covariant derivative and E is an endomorphism (a matrix in internal space). This generality is needed for the Standard Model (where A_μ is the gauge connection) and for induced gravity (where $g_{\mu\nu}$ is the spacetime metric).

The heat equation and the ansatz. The heat kernel $K(x, y; \tau) = \langle x | e^{-\tau D} | y \rangle$ satisfies:

$$\frac{\partial K}{\partial \tau} = -D_x K, \quad K(x, y; 0) = \frac{\delta^d(x - y)}{\sqrt{g}}. \quad (23.15)$$

For small τ and y near x , make the ansatz:

$$K(x, y; \tau) = \frac{1}{(4\pi\tau)^{d/2}} \Delta^{1/2}(x, y) e^{-\sigma(x, y)/(2\tau)} \sum_{n=0}^{\infty} a_n(x, y) \tau^n, \quad (23.16)$$

where $\sigma(x, y) = \frac{1}{2}d^2(x, y)$ is half the squared geodesic distance (Synge’s world function) and $\Delta(x, y) = \det(-\partial_\mu \partial'_\nu \sigma) / \sqrt{g(x)g(y)}$ is the Van Vleck–Morette determinant. The Gaussian factor $e^{-\sigma/(2\tau)}$ ensures the initial condition: as $\tau \rightarrow 0$, the kernel concentrates at $y = x$.

The recursion relation. Substituting the ansatz into the heat equation (23.15) and matching powers of τ gives a recursion:

$$(n+1) a_{n+1}(x, y) + \sigma^\mu \nabla_\mu a_{n+1}(x, y) = (\Delta^{-1/2} D_x \Delta^{1/2}) a_n(x, y), \quad (23.17)$$

where $\sigma^\mu = g^{\mu\nu} \partial_\nu \sigma$ is the tangent vector to the geodesic from y to x . The initial condition is $a_0(x, x) = \mathbf{1}$ (the identity in internal space).

At the coincidence point $y = x$ (where $\sigma = 0$, $\sigma^\mu = 0$, $\Delta = 1$), the recursion simplifies:

$$(n+1) a_{n+1}(x, x) = D_x a_n(x, x) \Big|_{y \rightarrow x}. \quad (23.18)$$

This determines $a_n(x, x)$ iteratively from $a_0(x, x) = \mathbf{1}$.

Explicit results for the coincidence-point coefficients. For the operator $D = -g^{\mu\nu} D_\mu D_\nu + E$ on a d -dimensional manifold with Riemann curvature $R_{\mu\nu\rho\sigma}$, Ricci tensor $R_{\mu\nu}$, scalar curvature R , and gauge field strength $\Omega_{\mu\nu} = [D_\mu, D_\nu] - \Gamma_{\mu\nu}^\rho D_\rho$:

$$a_0(x, x) = \text{tr } \mathbf{1}, \quad (23.19)$$

$$a_1(x, x) = \text{tr} \left(-E + \frac{1}{6} R \mathbf{1} \right), \quad (23.20)$$

$$a_2(x, x) = \text{tr} \left[\frac{1}{2} E^2 - \frac{1}{6} \nabla^2 E + \frac{1}{12} \Omega_{\mu\nu} \Omega^{\mu\nu} + \frac{1}{6} R E - \frac{1}{30} \nabla^2 R \mathbf{1} + \left(\frac{1}{72} R^2 - \frac{1}{180} R_{\mu\nu} R^{\mu\nu} + \frac{1}{180} R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma} \right) \mathbf{1} \right]. \quad (23.21)$$

These are the celebrated Seeley–DeWitt coefficients. Each is a local, gauge-invariant, diffeomorphism-invariant scalar built from the endomorphism E , the gauge field strength $\Omega_{\mu\nu}$, and the Riemann curvature. The traces are over the internal indices of the field (colour, flavour, spinor, etc.).

What each coefficient does.

a_0 : counts degrees of freedom. Inserted into the one-loop effective action $\Gamma^{(1)} = -\frac{1}{2} \int \frac{ds}{s} \frac{1}{(4\pi s)^{d/2}} \int d^d x \sum_n a_n s^n$, the a_0 term gives a quartic divergence $\propto 1/a^4$: the vacuum energy (cosmological constant).

a_1 : contains the endomorphism E and the scalar curvature R . *The E term* gives the quadratic divergence $\propto 1/a^2$ that renormalises the mass: $\delta m^2 \propto \text{tr}(E)/a^2$. For the Higgs field: $E = m_H^2 + \lambda \bar{\phi}^2 + y_t^2 |\bar{t}|^2 + \dots$, giving $\delta m_H^2 \propto (6\lambda + 6y_t^2 - \frac{9}{4}g^2 - \frac{3}{4}g'^2)/a^2$. *The R term* gives a quadratic divergence proportional to the scalar curvature: $\delta\Gamma \propto \text{tr}(\mathbf{1}) \cdot R/a^2$. This generates a gravitational kinetic term $\propto R/a^2$ in the effective action—**induced gravity**. Even if the tree-level action has no gravitational kinetic term ($G_N^{-1} = 0$ at tree level), the one-loop correction generates one:

$$\Gamma^{(1)} \supset \frac{1}{a^2} \cdot \frac{N_{\text{dof}}}{192\pi^2} \int d^4 x \sqrt{g} R, \quad (23.22)$$

where $N_{\text{dof}} = \sum_i \eta_i n_i$ is the weighted count of degrees of freedom. This is Sakharov’s induced gravity (1967): gravity as a one-loop effect of matter fields. The Newton constant is $G_N^{-1} \propto N_{\text{dof}}/a^2$ —finite on the lattice, a -dependent through the engine parameter.

a_2 : contains E^2 , $\Omega_{\mu\nu}^2$, and curvature-squared terms. *The E^2 term* renormalises the quartic coupling λ and the Yukawa couplings y_f : the beta functions of the SM scalar and Yukawa sectors. *The $\Omega_{\mu\nu}^2$ term* renormalises the gauge kinetic terms $F_{\mu\nu} F^{\mu\nu}$: the gauge beta functions (asymptotic freedom for $SU(3)_c$, asymptotic slavery for $U(1)_Y$). *The curvature-squared terms* (R^2 , $R_{\mu\nu} R^{\mu\nu}$, $R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma}$) give logarithmic corrections to the gravitational action: the running of higher-derivative gravity couplings. In $d = 4$, the Gauss–Bonnet combination $R_{\mu\nu\rho\sigma}^2 - 4R_{\mu\nu}^2 + R^2$ is topological and does not contribute to the equations of motion; the remaining two structures (R^2 and the Weyl tensor squared) run logarithmically. These are irrelevant operators (dimension 4 on a 4-manifold, but with two derivatives acting on the metric rather than on matter), and their effects are suppressed by $(E/M_P)^2$ at accessible energies.

23.11 Schwinger’s proper-time method

The propagator of the operator $D + m^2$ can be written as a Laplace transform of the heat kernel:

$$\frac{1}{D + m^2} = \int_0^\infty ds e^{-s(D+m^2)} = \int_0^\infty ds e^{-sm^2} e^{-sD}. \quad (23.23)$$

This is Schwinger’s proper-time representation (1951). The parameter s is the *proper time*—a Fock-layer bookkeeping parameter, not a physical time. Instead of inverting $D + m^2$ directly (which requires diagonalising D), Schwinger integrates the heat kernel e^{-sD} over proper time s weighted by e^{-sm^2} .

The proper-time integral converges at $s \rightarrow \infty$ (because e^{-sm^2} decays). The divergences come from $s \rightarrow 0$: the short proper-time region where the Seeley–DeWitt expansion (43.23) applies. The UV divergences of QFT are encoded in the small- s behaviour of the heat kernel.

The one-loop effective action. The one-loop contribution to the effective action (section 18.5) is:

$$\Gamma^{(1)} = \frac{1}{2} \ln \det D = \frac{1}{2} \text{Tr} \ln D = -\frac{1}{2} \int_0^\infty \frac{ds}{s} \text{Tr} e^{-sD}, \quad (23.24)$$

where Tr is the functional trace (trace over all indices and integration over all positions). This formula computes the one-loop effective action from the heat kernel alone: no Feynman diagrams, no momentum integrals, no combinatorics.

Substituting the Seeley–DeWitt expansion:

$$\Gamma^{(1)} = -\frac{1}{2} \int_0^\infty \frac{ds}{s} \frac{1}{(4\pi s)^{d/2}} \int d^d x \sum_{n=0}^\infty a_n(x) s^n. \quad (23.25)$$

The integral over s produces poles (in dimensional regularisation) or power/log divergences (with a cutoff) from the terms with $n \leq d/2$. For $d = 4$:

n	s -integral	Divergence	Physics
0	$\int ds s^{-3}$	Quartic: $\sim \Lambda^4$	Cosmological constant (vacuum energy)
1	$\int ds s^{-2}$	Quadratic: $\sim \Lambda^2$	Mass renormalisation
2	$\int ds s^{-1}$	Logarithmic: $\sim \ln \Lambda$	Coupling renormalisation, beta function
≥ 3	$\int ds s^{n-3}$	Finite	Higher-order corrections (irrelevant operators)

The UV divergent content of the one-loop effective action is entirely contained in a_0 , a_1 , and a_2 . These three coefficients determine the renormalisation of the cosmological constant, the mass, and the coupling—the same three renormalisations required by the Born output stability condition of section ??.

23.12 Beta functions from the heat kernel

The beta function for the coupling λ is extracted from the a_2 coefficient. For the ϕ^4 theory with $D = -\nabla^2 + m^2 + \frac{\lambda}{2}\bar{\phi}^2$ (the fluctuation operator around the background field $\bar{\phi}$):

$$a_2 = \frac{1}{2} \left(\frac{\lambda}{2} \bar{\phi}^2 \right)^2 + \dots = \frac{\lambda^2 \bar{\phi}^4}{8} + \dots \quad (23.26)$$

The $\bar{\phi}^4$ term in a_2 gives the logarithmically divergent correction to the ϕ^4 coupling. The stability condition $\mathcal{O}_{\text{Born}} = \mathcal{O}_{\text{phys}} + o(a/\lambda)$ requires absorbing this divergence into the running coupling $\lambda(a)$, giving:

$$\beta(\lambda) = \frac{d\lambda}{d \ln(1/a)} = \frac{3\lambda^2}{16\pi^2} + O(\lambda^3), \quad (23.27)$$

in agreement with the Hamiltonian-side Feshbach calculation (section 20.5) and the path-integral calculation.

For Yang–Mills gauge theory with $D = -D_\mu D^\mu$ acting on gauge fields, the a_2 coefficient contains the term $\frac{1}{12} \text{tr}(F_{\mu\nu} F^{\mu\nu})$, which gives the one-loop beta function:

$$\beta(g) = -\frac{11C_A}{48\pi^2} g^3 + O(g^5), \quad (23.28)$$

where C_A is the quadratic Casimir of the adjoint representation. The negative sign is asymptotic freedom: the coupling decreases at short distances. This result—the most important single number in quantum chromodynamics—drops out of the a_2 coefficient of the heat kernel, without a single Feynman diagram.

23.13 Anomalies and index theorems

The heat kernel provides a direct route to the chiral anomaly and the Atiyah–Singer index theorem.

The chiral anomaly. For a massless Dirac fermion coupled to a gauge field, the relevant operator is the squared Dirac operator $D = \not{D}^2 = -D_\mu D^\mu + \frac{1}{4}[D_\mu, D_\nu]\gamma^\mu\gamma^\nu$. The anomaly is the failure of the axial current $j_5^\mu = \bar{\psi}\gamma^\mu\gamma_5\psi$ to be conserved at the quantum level:

$$\partial_\mu j_5^\mu = 2 \lim_{s \rightarrow 0} \text{tr}(\gamma_5 K(x, x; s)) = \frac{1}{(4\pi)^{d/2}} \text{tr}(\gamma_5 a_{d/2}). \quad (23.29)$$

In $d = 4$: the anomaly is $\text{tr}(\gamma_5 a_2)$. From (23.14): $\text{tr}(\gamma_5 a_2) \propto \text{tr}(\gamma_5 F_{\mu\nu} F^{\mu\nu}) \propto F_{\mu\nu} \tilde{F}^{\mu\nu}$, giving the ABJ anomaly $\partial_\mu j_5^\mu = \frac{e^2}{16\pi^2} F_{\mu\nu} \tilde{F}^{\mu\nu}$.

The Atiyah–Singer index theorem. The index of the Dirac operator $\text{ind}(\not{D}) = n_+ - n_-$ (the difference in the number of left- and right-handed zero modes) is:

$$\text{ind}(\not{D}) = \lim_{s \rightarrow 0} \text{Tr}(\gamma_5 e^{-s\not{D}^2}) = \int d^d x \text{tr}(\gamma_5 a_{d/2}(x)). \quad (23.30)$$

This is a topological invariant: it depends only on the topology of the gauge field configuration (the instanton number), not on the detailed shape. The heat kernel provides a direct proof: the s dependence drops out because the non-zero modes cancel in pairs between $\gamma_5 = +1$ and $\gamma_5 = -1$ sectors, leaving only the zero modes whose count is topological.

23.14 The heat kernel as a unifying tool

The heat kernel $K(\tau) = e^{-\tau D}$ is a single operator-side object that unifies multiple computations:

Regime	What it gives	Where used
$\tau \rightarrow \infty$	Masses, gaps	QMC correlator decay (this section); spectral decomposition (section 18.6)
$\tau \rightarrow 0, a_0$	Vacuum energy (Λ^4)	Cosmological constant, Casimir energy
$\tau \rightarrow 0, a_1$	Mass renormalisation (Λ^2)	Hamiltonian renormalisation (section ??)
$\tau \rightarrow 0, a_2$	Beta functions ($\ln \Lambda$)	Coupling running, asymptotic freedom
$\tau \rightarrow 0, a_{d/2}$	Anomalies, index theorems	Chiral anomaly, instantons, topology
Proper-time integral	One-loop Γ	Effective potential, symmetry breaking

All of this is on the operator side. The heat kernel is a property of $e^{-\tau D}$ —a matrix exponential—not of a path-integral measure. Schwinger’s proper-time method is the technology; the Seeley–DeWitt expansion is the computational engine. No Feynman diagrams are needed at any stage.

This supports Schwinger’s original vision, which the textbook tradition has largely abandoned: the operator formalism is not merely an alternative to Feynman diagrams—it is a more systematic tool, one that extracts UV structure, anomalies, and index theorems from a single object ($e^{-\tau D}$) by examining its behaviour at different scales ($\tau \rightarrow \infty$ vs $\tau \rightarrow 0$).

23.15 The Fock-Born reading

Both the real-time and Euclidean path integrals are Fock-layer tools. Neither produces Born-layer observables directly. The bridges:

Masses: from the large- τ decay of $G_E^{(2)}(\tau)$ (Fock-layer) to the spectrometer reading (Born-layer).

Scattering: from the finite-volume energy shifts $\delta E_n(L)$ (Fock-layer, via QMC) through the Lüscher formula to scattering parameters (Born-layer).

Spectral functions: from $G_E^{(2)}(\tau)$ through inverse Laplace transform to $\rho(\mu)$ (Fock-layer), then to resonance peaks in $\sigma(E)$ (Born-layer).

Thermodynamics: from $\mathcal{Z}(\beta)$ (Fock-layer) to free energy, pressure, equation of state (Born-layer).

Beta functions and anomalies: from the Seeley–DeWitt coefficients a_n (Fock-layer, properties of the operator D) to the running couplings measured at different energy scales (Born-layer).

The heat kernel is the most powerful single tool in the Fock layer’s computational arsenal. It computes masses and anomalies from the same object, by examining different limits of the same operator exponential. The path integral is one method for evaluating it; the operator expansion is another. The Born-layer observables are the same either way.

24 Wilson Renormalization Group and Weinberg Effective Field Theory

The lattice engine at spacing a has parameters $\theta(a)$; the interface renders observables at resolution $\lambda \gg a$. The renormalization group is the framework that connects the two: it tracks how the engine parameters flow as the observation scale changes from a to λ , which interactions are generated, which are suppressed, and which survive to become the physics that the observer reports.

This section develops three perspectives — Wilson’s (start from the engine, integrate out), Weinberg’s (start from the symmetries, write down), and the Hamiltonian RG (Glazek–Wilson, Wegner: the same physics without passing through a path integral) — and synthesises them in the lattice-first framework.

We use continuum differential notation $(\partial_\mu \phi, \int d^d x)$ for readability. Every expression is defined on the lattice: ∂_μ is the lattice derivative, $\int d^d x = a^d \sum_x$, and momenta lie in the Brillouin zone $k_\mu \in (-\pi/a, \pi/a]$.

24.1 Dimensional analysis: one rule determines everything

The classification of operators into relevant, marginal, and irrelevant—which determines the entire structure of renormalization—follows from a single rule: **the action is dimensionless**.

In natural units ($\hbar = c = 1$), all quantities are measured in powers of mass M (equivalently, inverse length L^{-1} , equivalently energy E , equivalently inverse time T^{-1}). The *mass dimension* $[Q]$ of a quantity Q is the power of mass: $Q \sim M^{[Q]}$.

The rules

Rule 0 (the action is dimensionless): $[S] = 0$.

Since $S = \int d^d x \mathcal{L}$ and $[d^d x] = -d$ (each coordinate has $[x^\mu] = -1$, i.e. length):

$$[\mathcal{L}] = d. \quad (24.1)$$

Every term in the Lagrangian density must have mass dimension d . This single constraint determines the dimensions of all fields and couplings.

Rule 1 (derivatives add 1): $[\partial_\mu] = +1$.

Because $[\partial_\mu] = [1/x^\mu] = +1$.

Rule 2 (field dimensions from the kinetic term): The kinetic term determines $[\phi]$.

The kinetic term must have dimension d (to fit in $\mathcal{L}$). For each field type:

Real scalar field. Kinetic term: $\frac{1}{2}(\partial_\mu \phi)^2$. Dimension: $2[\partial] + 2[\phi] = 2 + 2[\phi] \stackrel{!}{=} d$, so:

$$\boxed{[\phi] = \frac{d-2}{2}}. \quad (24.2)$$

In $d = 4$: $[\phi] = 1$.

Dirac fermion. Kinetic term: $\bar{\psi} i \gamma^\mu \partial_\mu \psi$. Dimension: $2[\psi] + 1 \stackrel{!}{=} d$, so:

$$\boxed{[\psi] = \frac{d-1}{2}}. \quad (24.3)$$

In $d = 4$: $[\psi] = 3/2$.

Gauge field. Kinetic term: $-\frac{1}{4g^2}F_{\mu\nu}F^{\mu\nu} = -\frac{1}{4g^2}(\partial_\mu A_\nu - \partial_\nu A_\mu)^2$. If the gauge coupling g is extracted (so A_μ has its own dimension): $[F_{\mu\nu}] = [A] + 1$, and $[F^2] = 2[A] + 2 \stackrel{!}{=} d$, so:

$$\boxed{[A_\mu] = \frac{d-2}{2}}. \quad (24.4)$$

In $d = 4$: $[A_\mu] = 1$. (Same as the scalar. The gauge field and the scalar have the same mass dimension in any d .)

Metric. The metric $g_{\mu\nu}$ is dimensionless:

$$\boxed{[g_{\mu\nu}] = 0}. \quad (24.5)$$

This is because $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$ has $[ds^2] = 2[dx] = -2$, which requires $[g_{\mu\nu}] + 2[dx] = -2$, so $[g_{\mu\nu}] = 0$. The metric is a ratio of lengths, not a length.

Rule 3 (coupling dimensions from the interaction): The coupling makes up the difference between the operator dimension and d .

For an interaction $g_i\mathcal{O}_i$ in the Lagrangian density: $[g_i] + [\mathcal{O}_i] \stackrel{!}{=} d$, so:

$$\boxed{[g_i] = d - [\mathcal{O}_i]}. \quad (24.6)$$

This determines the classification:

$$[g_i] > 0 \quad \Rightarrow \quad [\mathcal{O}_i] < d \quad \Rightarrow \quad \text{relevant}, \quad (24.7)$$

$$[g_i] = 0 \quad \Rightarrow \quad [\mathcal{O}_i] = d \quad \Rightarrow \quad \text{marginal}, \quad (24.8)$$

$$[g_i] < 0 \quad \Rightarrow \quad [\mathcal{O}_i] > d \quad \Rightarrow \quad \text{irrelevant}. \quad (24.9)$$

The key examples

Operator ($d = 4$)	$[\mathcal{O}]$	$[g]$	Class	a -scaling
$\mathbf{1}$ (cosmological const.)	0	4	Relevant	$g \sim 1/a^4$
ϕ^2 (mass)	2	2	Relevant	$g \sim 1/a^2$
$\bar{\psi}\psi$ (fermion mass)	3	1	Relevant	$g \sim 1/a$
$(\partial\phi)^2$ (scalar kinetic)	4	0	Marginal	$g \sim \ln(1/a)$
ϕ^4 (quartic)	4	0	Marginal	$g \sim \ln(1/a)$
$F_{\mu\nu}F^{\mu\nu}$ (gauge kinetic)	4	0	Marginal	$g \sim \ln(1/a)$
$\bar{\psi}\gamma^\mu\psi A_\mu$ (gauge-fermion)	4	0	Marginal	$g \sim \ln(1/a)$
$y\bar{\psi}\phi\psi$ (Yukawa)	4	0	Marginal	$g \sim \ln(1/a)$
R (Ricci scalar)	2	2	Relevant	$g \sim 1/a^2$
R^2 (curvature squared)	4	0	Marginal	$g \sim \ln(1/a)$
ϕ^6	6	-2	Irrelevant	$g \sim a^2$
R^3	6	-2	Irrelevant	$g \sim a^2$
F^4	8	-4	Irrelevant	$g \sim a^4$

Reading the table. Every operator of the Standard Model $((\partial\phi)^2, \phi^4, F^2, \bar{\psi}\gamma\partial\psi, y\bar{\psi}\phi\psi, \lambda\phi^4)$ has dimension $\leq 4 = d$: they are all relevant or marginal. This is *renormalisability*: the SM Lagrangian contains only operators whose couplings have non-negative mass dimension. Every higher-dimension operator $(\phi^6, F^4, R^3, \dots)$ is irrelevant: its effect on Born-layer observables is suppressed by $(a/\lambda)^{[\mathcal{O}]-d}$ and absorbed into the $o(a/\lambda)$ correction of rendering stability.

Why $[g_{\mu\nu}] = 0$ changes everything. The table reveals why gravity is structurally different from gauge theory. The gauge kinetic term $F_{\mu\nu}F^{\mu\nu}$ has dimension $4 = d$: marginal. The gravitational kinetic term R has dimension $2 < d$: relevant. This mismatch arises because $[A_\mu] = 1$ but $[g_{\mu\nu}] = 0$:

$$[F^2] = 2([A] + 1) = 2 \cdot 2 = 4 = d, \quad [R] = [g] + 2 = 0 + 2 = 2 < d. \quad (24.10)$$

The metric’s vanishing dimension pushes the gravitational operators two units lower than the corresponding gauge operators, promoting the first correction (R^2) from irrelevant to marginal and the kinetic term (R) from marginal to relevant.

This is the dimensional-analysis origin of the structural comparison developed in Section ??: the $\langle JJ \rangle$ correlator generates one marginal operator and nothing relevant (because the Ward identity kills $m_A^2 A^2$); the $\langle TT \rangle$ correlator generates two relevant and two marginal operators (because no Ward identity kills Λ or $M_P^2 R$).

In ML terms. Dimensional analysis is feature scaling. The “mass dimension” of an operator is like the number of pixels that the corresponding feature spans in the input image: a feature that spans the whole image (low dimension, relevant) dominates the prediction at any resolution; a feature that spans a single pixel (high dimension, irrelevant) washes out when the image is downsampled. The relevant/marginal/irrelevant classification is the QFT analogue of deciding which features survive coarse-graining—which is exactly the question that Wilson’s RG answers.

24.2 From dimensional analysis to renormalizability: the lattice-first argument

We now connect dimensional analysis to the rendering-stability framework developed in Sections 20.3 and 21.9. The argument is:

Dimensional analysis classifies operators.

The classification determines which operators need engine parameters.

The count of needed parameters determines renormalizability.

Each step is a theorem within the lattice-first framework.

Step 1: On the lattice, all operators are present

Section 24.7 established: the Wilson RG (or equivalently, the Feshbach formula) generates every operator compatible with the symmetries. At lattice spacing a , the effective action contains:

$$S_a[\phi] = \sum_i g_i(a) \int d^d x \mathcal{O}_i(\phi), \quad (24.11)$$

where the sum runs over *all* local operators $\mathcal{O}_i$ consistent with the symmetries of the theory ($\mathbb{Z}_2$ for the scalar, $SU(3) \times SU(2) \times U(1)$ BRST for the SM, etc.). The couplings $g_i(a)$ depend on the lattice spacing. Some were present in the original action; others were induced by the RG. All are nonzero in general.

On the lattice, this sum is finite: the lattice has finitely many sites and finitely many independent operators (a polynomial in ϕ_n and its finite differences, up to some maximum degree set by the

lattice geometry). In the differential notation, the sum is formally infinite but every term is a well-defined lattice expression.

Step 2: Dimensional analysis determines the Born-layer impact

Each operator $\mathcal{O}_i$ has mass dimension $[\mathcal{O}_i] = \Delta_i$, and its coupling has dimension $[g_i] = d - \Delta_i$ (Section 24.1). The contribution of $\mathcal{O}_i$ to a Born-layer observable $\mathcal{O}_{\text{Born}}$ at observation scale λ can be estimated by dimensional analysis alone.

The operator enters the Born-layer observable through loop corrections, Feynman diagrams, or non-perturbative effects. Regardless of the method, the contribution scales as:

$$\delta\mathcal{O}_{\text{Born}}|_{\mathcal{O}_i} \sim g_i(a) \cdot \lambda^{-[g_i]} \sim g_i(a) \cdot \lambda^{\Delta_i-d}. \quad (24.12)$$

This is the key formula. The factor λ^{Δ_i-d} converts the coupling (which has dimension $d - \Delta_i$) into a dimensionless contribution to the observable. On the lattice, $g_i(a) \sim a^{-(d-\Delta_i)}$ (the natural scale of a coupling of dimension $d - \Delta_i$ on a lattice of spacing a), giving:

$$\delta\mathcal{O}_{\text{Born}}|_{\mathcal{O}_i} \sim \frac{1}{a^{d-\Delta_i}} \cdot \frac{1}{\lambda^{\Delta_i-d}} = \left(\frac{a}{\lambda}\right)^{\Delta_i-d}. \quad (24.13)$$

Now read off the three cases:

Irrelevant ($\Delta_i > d$): $\delta\mathcal{O} \sim (a/\lambda)^{\Delta_i-d} \rightarrow 0$ as $a/\lambda \rightarrow 0$.

The contribution vanishes in the rendering-stability limit. It is automatically $o(a/\lambda)$. No engine parameter is needed to absorb it — it is a lattice artefact that dies on its own.

Marginal ($\Delta_i = d$): $\delta\mathcal{O} \sim (a/\lambda)^0 = O(1)$, with logarithmic corrections $\sim \ln(\lambda/a)$.

The contribution does NOT vanish. It produces an $O(1)$ correction (up to logs) to the Born-layer observable. If $g_i(a)$ is not adjusted as a changes, the Born output changes by $O(\ln(\lambda/a))$. An engine parameter must be assigned and tuned to maintain rendering stability.

Relevant ($\Delta_i < d$): $\delta\mathcal{O} \sim (\lambda/a)^{d-\Delta_i} \rightarrow \infty$ as $a/\lambda \rightarrow 0$.

The contribution grows without bound. It dominates the Born-layer observable. The engine parameter $g_i(a)$ must be adjusted precisely (it must scale as $1/a^{d-\Delta_i}$) to produce a finite Born-layer output. The required precision is set by the condition number κ (Section ??).

Summary:

Class	Δ_i vs d	Born impact	Engine parameter?	a -scaling of g_i
Relevant	$< d$	$(\lambda/a)^{d-\Delta_i}$: grows	Yes (must tune)	$g_i \sim 1/a^{d-\Delta_i}$
Marginal	$= d$	$O(\ln \lambda/a)$: persists	Yes (must tune)	$g_i \sim \ln(1/a)$
Irrelevant	$> d$	$(a/\lambda)^{\Delta_i-d}$: dies	No (auto-absorbed)	$g_i \sim a^{\Delta_i-d}$

Step 3: Renormalizability = finite count of relevant + marginal operators

A theory is **renormalisable** if the number of relevant and marginal operators (those with $\Delta_i \leq d$) is finite. In this case: finitely many engine parameters $\theta(a)$ suffice for the rendering-stability condition $\mathcal{O}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$. All other operators are irrelevant: their contributions are automatically $o(a/\lambda)$ and need not be separately tuned.

The question: when is the count finite?

The counting argument in $d = 4$. An operator $\mathcal{O}_i$ built from n_ϕ scalars, n_ψ fermions, n_A gauge fields, and n_∂ derivatives has dimension:

$$\Delta_i = n_\phi \cdot 1 + n_\psi \cdot \frac{3}{2} + n_A \cdot 1 + n_\partial \cdot 1 \quad (d = 4). \quad (24.14)$$

(Using $[\phi] = [A_\mu] = 1$ and $[\psi] = 3/2$ in $d = 4$.) The condition $\Delta_i \leq 4$ constrains:

$$n_\phi + n_A + n_\partial + \frac{3}{2}n_\psi \leq 4. \quad (24.15)$$

This has only finitely many non-negative integer solutions. The complete list (for a single real scalar, omitting total derivatives):

Operator	Δ	$[g]$	Class
1 (vacuum energy)	0	4	Relevant
ϕ^2	2	2	Relevant
ϕ^3	3	1	Relevant
ϕ^4	4	0	Marginal
$(\partial_\mu \phi)^2$	4	0	Marginal

Five operators. But ϕ^3 is forbidden by $\mathbb{Z}_2$ symmetry ($\phi \rightarrow -\phi$), and the vacuum energy **1** does not affect scattering (it is a constant shift of H). For the ϕ^4 theory with $\mathbb{Z}_2$: **three engine parameters** (m^2, λ, Z) suffice. The theory is renormalisable.

For any $d = 4$ theory with a finite number of field species: the constraint (24.15) has finitely many solutions, so the number of relevant+marginal operators is finite, and the theory is renormalisable.

For the Standard Model: the constraint, combined with Lorentz invariance and BRST (Section 39.2), gives exactly 19 independent operators of dimension ≤ 4 . Nineteen engine parameters suffice for rendering stability.

Why $d = 4$ is special. The counting depends on d through the field dimensions:

d	$[\phi]$	$[\psi]$	Max. monomial degree for $\Delta \leq d$
2	0	1/2	∞ (any power of ϕ)
3	1/2	1	$\phi^6, \phi^4 \partial, \dots$
4	1	3/2	$\phi^4, \bar{\psi} \psi \phi, F^2$
5	3/2	2	$\phi^2, \bar{\psi} \psi$ only
6	2	5/2	ϕ^2 only

In $d = 2$: every power of ϕ is relevant (since $[\phi] = 0$), and infinitely many operators have $\Delta \leq 2$. Two-dimensional field theories require special treatment (conformal field theory).

In $d = 4$: the list is finite and short. The quartic coupling ϕ^4 is exactly marginal ($\Delta = 4 = d$). Quintic and higher are irrelevant. This is why $d = 4$ admits renormalisable interacting theories with finitely many parameters.

In $d > 4$: even the quartic coupling ϕ^4 has $\Delta = 2d/2 + 2 > d$ (for $d > 4$: $\Delta = 4 > d$ only if $d < 4$). Actually, $[\phi^4] = 4[\phi] = 4(d-2)/2 = 2(d-2)$. For $d = 5$: $[\phi^4] = 6 > 5$, so ϕ^4 is irrelevant. No non-trivial self-interaction is relevant or marginal for scalars in $d > 4$. Interacting scalar field theories are non-renormalisable in $d > 4$.

Step 4: The lattice-first punchline

Now state the result in the lattice-first language.

Theorem (renormalizability from dimensional analysis). *For a d -dimensional QFT with a finite number of field species, the rendering-stability condition $\mathcal{O}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$ can be satisfied with finitely many engine parameters $\theta(a)$ if and only if the number of independent operators with $\Delta_i \leq d$ (after imposing all symmetry constraints) is finite.*

Proof sketch.

“If” direction: if $N_{\text{rel+marg}}$ operators have $\Delta_i \leq d$, assign each an engine parameter $g_i(a)$. The rendering-stability condition gives $N_{\text{rel+marg}}$ equations for $N_{\text{rel+marg}}$ unknowns. The solution exists and is unique (by local invertibility, Section 24.6). All remaining operators have $\Delta_i > d$ and contribute $o(a/\lambda)$ automatically. Therefore $N_{\text{rel+marg}}$ parameters suffice.

“Only if” direction: if infinitely many operators have $\Delta_i \leq d$, each contributes $O(1)$ or larger to the Born-layer output and must be separately tuned. Infinitely many engine parameters are needed. The theory is non-renormalisable.

At finite a : the distinction softens. On the lattice at finite a , there are finitely many lattice sites and finitely many independent operators. Every theory is “trivially renormalisable” at finite a : finitely many parameters, finitely many rendering conditions, finitely many everything. The distinction between renormalisable and non-renormalisable sharpens only in the limit $a \rightarrow 0$, where the number of independent operators grows without bound (the lattice’s mode count increases as $(1/a)^d$).

In the lattice-first framework: renormalisability is a property of the *continuum extrapolation*, not of the engine at finite a . A non-renormalisable theory (like gravity) is perfectly well-defined on the lattice at every a . The “non-renormalisability” means: as $a \rightarrow 0$, the number of engine parameters needed to maintain rendering stability grows without bound. At finite a , the count is finite and the engine runs.

This connects to the type-B classification of Section ??: non-renormalisability is a question about the rendering’s infinite-resolution limit, not about the physics at finite a .

The complete logical chain

Step	Content
1	$[S] = 0$ The action is dimensionless $\Rightarrow$ field dimensions $[\phi]$, $[\psi]$, $[A]$, $[g]$ are determined
2	$[g_i] = d - \Delta_i$ Coupling dimension determines relevant/marginal/irrelevant classification
3	Wilson / Feshbach All operators are generated; irrelevant ones are suppressed by $(a/\lambda)^{\Delta_i - d}$
4	Rendering stability $\mathcal{O}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda) \Rightarrow$ only relevant + marginal operators need engine parameters
5	Counting $n_\phi + n_A + n_\partial + \frac{3}{2}n_\psi \leq d$ has finitely many solutions $\Rightarrow$ finitely many parameters $\Rightarrow$ renormalisable
6	BRST Gauge symmetry reduces the count further (hundreds $\rightarrow$ 19 for the SM)

This is the complete logical chain from $[S] = 0$ (one axiom) to “the Standard Model is renormal-

isable with 19 engine parameters” (the final result). Every step is a computation within the lattice-first framework. No continuum limit is taken. No infinity is encountered. No infinity is “removed.”

24.3 Wilson’s renormalization group: the flow of couplings

The coupling space

The most general action for a real scalar field with $\mathbb{Z}_2$ symmetry ($\phi \rightarrow -\phi$) contains infinitely many operators:

$$S[\phi] = \int d^d x \left[\frac{1}{2}(\partial_\mu \phi)^2 + \frac{r}{2}\phi^2 + \frac{u_4}{4!}\phi^4 + \frac{u_6}{6!}\phi^6 + \frac{u_8}{8!}\phi^8 + c_1(\partial^2 \phi)^2 + c_2\phi^2(\partial_\mu \phi)^2 + \dots \right]. \quad (24.16)$$

Each operator has a coupling constant. The *coupling space* is the infinite-dimensional space $\theta = (r, u_4, u_6, u_8, c_1, c_2, \dots)$. A point θ specifies a particular engine.

The textbook ϕ^4 theory sits on a low-dimensional subspace: $r \neq 0$, $u_4 \neq 0$, everything else zero. This is a specific choice, not a general requirement.

The RG transformation

Divide the field into slow and fast modes: $\phi = \phi^< + \phi^>$, where $\phi^<$ has momenta $|k| < \Lambda/b$ and $\phi^>$ has momenta $\Lambda/b < |k| < \Lambda$ (with $\Lambda = \pi/a$, $b > 1$). Integrate out the fast modes:

$$e^{-S_{\text{eff}}[\phi^<]} = \int \mathcal{D}\phi^> e^{-S[\phi^< + \phi^>]}. \quad (24.17)$$

Rescale momenta ($k \rightarrow bk$) and fields ($\phi^< \rightarrow b^{(d+2-\eta)/2}\phi'$) to restore the original cutoff. The combined operation is the RG map: $\theta(a) \xrightarrow{\mathcal{R}} \theta'(ba)$.

The central fact: $\mathcal{R}$ generates new interactions

Even if the initial action has $u_6 = u_8 = c_1 = c_2 = \dots = 0$ (only r and u_4 nonzero), a single RG step generates *all* operators compatible with the $\mathbb{Z}_2$ symmetry, with calculable coefficients.

The mechanism is explicit. The fast-mode integral over $\phi^>$ generates corrections to the slow-mode action through contractions of $\phi^>$ inside $S[\phi^< + \phi^>]$:

A mass correction (one contraction):

$$\delta r = \frac{u_4}{2} \int_{\Lambda/b}^{\Lambda} \frac{d^d k}{(2\pi)^d} \frac{1}{k^2 + r} \equiv \frac{u_4}{2} I_1. \quad (24.18)$$

A renormalized quartic (one-loop “fish” diagram):

$$\delta u_4 = -\frac{3u_4^2}{2} \int_{\Lambda/b}^{\Lambda} \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2 + r)^2} \equiv -\frac{3u_4^2}{2} I_2. \quad (24.19)$$

An induced sextic (generated from zero):

$$u'_6 = c_6^{(1)} u_4^2 I_2 + c_6^{(2)} u_4^3 I_3 + \dots \quad (24.20)$$

Two or more u_4 vertices connected by fast propagators, with six external $\phi^<$ legs. Starting from $u_6(a) = 0$: after one step, $u_6(ba) \neq 0$.

An induced octic: $u'_8 \sim u_4^3 I_3 + u_4^4 I_4 + \dots$

Induced derivative interactions: $c'_1 \sim u_4 J_1 + u_4^2 J_2 + \dots$, from the k -dependence of the loop integrals (J_n are derivatives of I_n with respect to external momenta).

Every operator compatible with $\mathbb{Z}_2$ is generated, with a coefficient that is a polynomial in u_4 times momentum-shell integrals.

The generated couplings are irrelevant. Although the RG generates $u_6, u_8, c_1, \dots$ from the quartic seed, these are all irrelevant operators (scaling dimension $d_i > d$). Under subsequent RG steps, they shrink: $u_{2n}(\lambda) \sim u_4(a)^n \cdot (a/\lambda)^{d_{2n}-d} \rightarrow 0$ as $\lambda/a \rightarrow \infty$. The RG generates them and then suppresses them. At the rendering scale $\lambda \gg a$: they are present but small, proportional to powers of (a/λ) .

The beta functions

The infinitesimal RG ($b = 1 + d\ell$) gives the beta functions:

$$\beta_r \equiv \frac{dr}{d\ell} = 2r + \frac{u_4 \Omega_d \Lambda^d}{2(2\pi)^d (\Lambda^2 + r)}, \quad (24.21)$$

$$\beta_{u_4} \equiv \frac{du_4}{d\ell} = (4 - d) u_4 - \frac{3u_4^2 \Omega_d \Lambda^d}{2(2\pi)^d (\Lambda^2 + r)^2}, \quad (24.22)$$

where $\Omega_d = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of the unit sphere.

Fixed points and critical exponents

The Gaussian fixed point ($r^* = 0, u_4^* = 0$): the free theory. Eigenvalues of the linearised flow: $\theta_r = 2$ (relevant), $\theta_{u_4} = 4 - d$ (relevant for $d < 4$, marginal for $d = 4$, irrelevant for $d > 4$), $\theta_{u_6} = 6 - 2d$ (irrelevant for $d > 3$).

The Wilson–Fisher fixed point ($d = 4 - \epsilon, \epsilon > 0$):

$$u_4^* = \frac{16\pi^2}{3} \epsilon + O(\epsilon^2), \quad r^* = -\frac{u_4^* I_1}{2(2 - \epsilon)} + O(\epsilon^2). \quad (24.23)$$

Critical exponents at the Wilson–Fisher fixed point:

$$\nu = \frac{1}{2} + \frac{\epsilon}{12} + O(\epsilon^2) \quad (\xi \sim |T - T_c|^{-\nu}), \quad (24.24)$$

$$\eta = O(\epsilon^2) \quad (\text{anomalous dimension}), \quad (24.25)$$

$$\gamma = 1 + \frac{\epsilon}{6} + O(\epsilon^2) \quad (\text{susceptibility}). \quad (24.26)$$

At $d = 3$ ($\epsilon = 1$): this fixed point governs the 3D Ising universality class (the liquid-gas critical point, the ferromagnet, the superfluid transition). The exponents are universal: they depend only on the symmetry ($\mathbb{Z}_2$) and the dimensionality ($d = 3$), not on the lattice details.

Triviality at $d = 4$. At $\epsilon = 0$: $u_4^* \rightarrow 0$. The quartic coupling is marginally irrelevant: $du_4/d\ell = -3u_4^2/(16\pi^2) < 0$. Solution:

$$u_4(\ell) = \frac{u_4(0)}{1 + \frac{3u_4(0)}{16\pi^2} \ell}. \quad (24.27)$$

The coupling flows to zero logarithmically. In the continuum limit ($\ell \rightarrow \infty$): the theory is free. At finite lattice spacing (ℓ finite): the theory is interacting. Triviality is a property of the extrapolation, not of the engine.

Relevant, marginal, and irrelevant

At a fixed point: the linearised RG classifies all operators:

Type	θ	Under coarse-graining	Role at $\lambda \gg a$
Relevant	> 0	Grows: $\sim (\lambda/a)^\theta$	Free parameter (must be measured)
Marginal	$= 0$	Unchanged (log corrections)	Logarithmic running
Irrelevant	< 0	Shrinks: $\sim (a/\lambda)^{ \theta }$	Lattice artefact (suppressed)

The irrelevant operators encode the engine's microscopic structure and are suppressed at the rendering scale by powers of a/λ . The rendering-stability condition $\mathcal{O}_{\text{phys}} = f(\theta) + o(a/\lambda)$ is the statement that only relevant and marginal couplings matter at long distances.

24.4 Hamiltonian renormalization

Wilson's RG operates on the Euclidean action in the path-integral formalism. The Hamiltonian RG operates directly on H in the operator formalism, without passing through a path integral. This is the RG native to the Fock-layer engine.

The Glazek–Wilson similarity renormalization group

Glazek and Wilson (1993–94) proposed a continuous family of unitary transformations that decouple high-energy from low-energy degrees of freedom in Fock space.

The setup. The Hamiltonian H in a Fock basis has matrix elements H_{mn} between states of energies E_m, E_n . Off-diagonal elements with $|E_m - E_n| \gg \Lambda_s$ couple widely separated energy scales. The goal: eliminate these elements systematically, reducing H to *band-diagonal* form where only states within an energy band Λ_s are coupled.

The flow equation. A one-parameter family of unitarily equivalent Hamiltonians:

$$H_s = U_s H U_s^\dagger, \quad U_s = \mathcal{T} \exp\left(-\int_0^s \eta_{s'} ds'\right), \quad (24.28)$$

where $\eta_s = -\eta_s^\dagger$ is the anti-Hermitian generator and s has dimensions of inverse energy squared ($\Lambda_s \sim 1/\sqrt{s}$). The flow:

$$\boxed{\frac{dH_s}{ds} = [\eta_s, H_s]}. \quad (24.29)$$

The generator η_s is chosen to suppress off-diagonal elements beyond bandwidth Λ_s . Glazek and Wilson's choice:

$$(\eta_s)_{mn} = \text{sgn}(E_m^d - E_n^d) \cdot (H_s)_{mn} \cdot f_s(E_m^d, E_n^d), \quad (24.30)$$

where E_m^d is the diagonal (free) energy and f_s is a smooth cutoff: $f_s \rightarrow 0$ when $|E_m^d - E_n^d| < \Lambda_s$, $f_s \rightarrow 1$ when $|E_m^d - E_n^d| \gg \Lambda_s$.

As s increases (Λ_s decreases): off-diagonal elements above Λ_s are suppressed; H_s becomes band-diagonal. At $s \rightarrow \infty$ ($\Lambda_s \rightarrow 0$): H_∞ is diagonal (if the spectrum is non-degenerate).

Wegner's flow equation

Wegner (1994) independently proposed an algebraic choice of generator:

$$\eta_s = [H_s^d, H_s], \quad (24.31)$$

where H_s^d is the diagonal part of H_s . The flow:

$$\frac{dH_s}{ds} = [[H_s^d, H_s], H_s]. \quad (24.32)$$

This choice has a virtue: the off-diagonal norm decreases monotonically:

$$\frac{d}{ds} \text{Tr}(H_s^{od})^2 = -2 \text{Tr}[\eta_s, H_s^d]^2 \leq 0. \quad (24.33)$$

The flow automatically diagonalises H : the off-diagonal part shrinks at every step.

What the flow generates

The Hamiltonian RG generates effective interactions in Fock space, just as Wilson’s path-integral RG generates effective operators in the action.

Start from a Hamiltonian with a kinetic term H_0 and a simple interaction V (e.g., a ϕ^4 vertex). The flow equation (24.29) generates:

At order V : the diagonal part of V is unchanged (it was already within the band).

At order V^2 : the commutator $[\eta_s, V]$ generates new interactions — the Fock-space analogue of the induced u_6 in Wilson’s RG. Explicitly: a ϕ^4 vertex connects a two-particle state to a four-particle state; the flow generates an effective six-particle interaction (two V ’s connected by a high-energy propagator that is being eliminated).

At order V^3 : eight-particle interactions, derivative couplings, etc.

The pattern is identical to Wilson’s: the flow fills in the complete tower of Fock-space interactions compatible with the symmetries, starting from a single seed interaction V . The generated interactions are suppressed by powers of V/Λ_s (the ratio of the interaction strength to the bandwidth): they are the Hamiltonian analogue of the irrelevant operators.

The correspondence: path integral and Hamiltonian

	Wilson (path integral)	Glazek–Wilson / Wegner
Object	Euclidean action $S_E[\phi]$	Hamiltonian H
Operates on	Field configurations	Fock-space states
“Integrate out”	Momentum shell $\Lambda/b < k < \Lambda$	Off-diagonal $ E_m - E_n > \Lambda_s$
Output	Effective action $S'_E[\phi^<]$	Band-diagonal H_s
Flow parameter	$\ell = \ln b$ (length scale)	s (inverse energy ²)
Generates	$u_6, u_8, c_1, \dots$ in S	6-body, 8-body, $\dots$ in H
Fixed point	$\mathcal{R}[S^*] = S^*$	$[H_s^d, H_s] = 0$ (diagonal)
Sign problem?	Yes (at finite density)	No (unitary transformation)
Real-time dynamics?	No (Euclidean signature)	Yes ($e^{-iH_s t}$ directly)

The two frameworks are complementary. Wilson’s is more developed for equilibrium critical phenomena (Monte Carlo). The Hamiltonian RG is essential for real-time dynamics, finite-density physics, and quantum simulation.

Applications

Nuclear physics: the SRG (Bogner–Schwenk–Furnstahl) is now standard for softening the nucleon-nucleon potential. The bare V_{NN} from chiral EFT has strong short-range repulsion

coupling low- and high-momentum states. The flow $V_s = U_s V U_s^\dagger$ decouples the high-momentum modes, producing a “soft” potential that converges rapidly in many-body calculations.

Light-front QCD: the Glazek–Wilson SRG on the light front (where the vacuum is trivial and the Fock structure is simpler) computes meson and baryon spectra, parton distributions, and form factors by systematically eliminating high-energy Fock sectors.

Condensed matter: Wegner’s flow for the Kondo model systematically decouples the impurity spin from high-energy conduction electrons, generating the Kondo temperature $T_K \sim D e^{-1/(J\rho_F)}$ as a non-perturbative output of the flow.

24.5 The lattice-first synthesis

The engine at spacing a

The engine is a lattice Hamiltonian $H(a)$ with couplings $\theta(a)$. At the engine scale: all couplings are $O(1)$ in lattice units. The irrelevant couplings are not small at the engine scale; they are suppressed only at the rendering scale.

The rendering at scale λ

Apply the RG (Wilson or Glazek–Wilson/Wegner) from a to λ . The effective couplings:

Relevant/marginal: controlled by $\theta(a)$ through the flow. These carry the physical information from engine to interface.

Irrelevant: suppressed by $(a/\lambda)^{|\theta|}$. These are the lattice artefacts — the engine’s fingerprints, invisible at the rendering scale.

Generated operators: u_6, u_8 , derivative couplings, Lorentz-violating terms — all generated by the RG from the seed couplings at a , all irrelevant, all suppressed. The rendering is cleaner than the engine: the RG washes out the microscopic details.

The rendering-stability condition:

$$\mathcal{O}_{\text{phys}}(\lambda) = \mathcal{O}_{\text{phys}}(\infty) + O((a/\lambda)^2). \quad (24.34)$$

What Weinberg sees

At the rendering scale: the observer sees an effective theory with emergent symmetries imposed, relevant/marginal operators parametrised, and irrelevant operators discarded. Weinberg measures the couplings and predicts to any desired order in E/Λ .

Weinberg’s success (the predictive power of the EFT) confirms the lattice-first picture: the rendering is universal. Weinberg’s limitation (he cannot derive the couplings, only measure them) is the limitation of the rendering: the interface does not reveal the engine.

The Hamiltonian advantage

The Glazek–Wilson/Wegner flow has structural advantages for the lattice-first programme:

No sign problem. At finite baryon density: the Euclidean fermion determinant becomes complex, Monte Carlo fails, and Wilson’s path-integral RG cannot be implemented. The Hamiltonian RG has no sign problem: H is Hermitian and U_s is unitary regardless of the chemical potential.

Real-time dynamics. The effective Hamiltonian H_s generates real-time evolution $e^{-iH_s t}$ directly. Spectral functions, transport coefficients, and non-equilibrium processes are accessible without the ill-posed analytic continuation from Euclidean to Minkowski.

The native language of the engine. The four-line source code ((46.10)–(46.13)) is Hamiltonian: $h_{t+dt} = e^{-iHdt}h_t$. The Glazek–Wilson/Wegner flow is the Hamiltonian RG that connects the engine’s microscopic $H(a)$ to its macroscopic predictions — the RG that operates directly on the engine, without passing through the rendering.

The final comparison

	Wilson	Weinberg	Hamiltonian RG	Lattice-first
Starts from	$S_E(a)$	Symmetries	$H(a)$	$H(a)$ on lattice
Direction	Top-down	Bottom-up	Top-down	Top-down
Knows UV?	Yes	No	Yes	Yes (IS the UV)
Symmetry	All	Assumed exact	All	Exact + emergent
Framework	Path integral	Path integral	Operator	Operator
Sign problem	Yes	Yes	No	No

24.6 Equivalence of the lattice-first and Wilson views

Two views of renormalization have been developed in parallel:

The lattice-first view. The engine at lattice spacing a has parameters $\theta(a)$. The rendering-stability condition requires that all Born-layer observables are a -independent:

$$\mathcal{O}_{\text{Born}}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda). \quad (\text{L})$$

This view does not specify *how* $\theta(a)$ is constructed; it only specifies *what* it must satisfy.

The Wilson view. Start at lattice spacing a with parameters $\theta(a)$. Integrate out the momentum shell $\pi/a' < |\mathbf{k}| < \pi/a$ (where $a' > a$) to obtain an effective theory at the coarser lattice spacing a' with shifted parameters. The Wilson RG map:

$$\mathcal{R}_{a \rightarrow a'} : \theta(a) \mapsto \theta(a') = \mathcal{R}_{a \rightarrow a'}[\theta(a)]. \quad (\text{W})$$

This view specifies *how* $\theta(a')$ is constructed from $\theta(a)$; it does not mention Born-layer observables or rendering stability.

The two views appear to be different prescriptions. We now prove they are mathematically equivalent: the Wilson RG map generates the unique family $\theta(a)$ satisfying rendering stability, and conversely, any rendering-stable family must satisfy the Wilson RG equation.

Theorem (Equivalence)

Let $\mathcal{R}_{a \rightarrow a'}$ denote the exact Wilson RG map (the Feshbach–Löwdin mode elimination from cutoff π/a to π/a'). Let $\mathcal{O}[\theta, a]$ denote a Born-layer observable computed from the engine with parameters θ at lattice spacing a . Then:

(A) *Wilson $\Rightarrow$ rendering stability:* $\mathcal{O}[\mathcal{R}_{a \rightarrow a'}(\theta(a)), a'] = \mathcal{O}[\theta(a), a]$ exactly.

(B) *Rendering stability $\Rightarrow$ Wilson:* if $\theta(a)$ is a family satisfying (L) and the map $\theta \mapsto \mathcal{O}[\theta, a']$ is locally invertible, then $\theta(a') = \mathcal{R}_{a \rightarrow a'}[\theta(a)] + o(a'/\lambda)$.

In short: Wilson generates rendering stability exactly, and rendering stability implies Wilson up to lattice artefacts.

Proof of (A): Wilson preserves observables exactly

The Feshbach formula (Section 20.5) gives the exact effective Hamiltonian on the coarser Fock space $\mathcal{F}_{a'} = P\mathcal{F}_a$:

$$H_{\text{eff}}(E) = H_{PP} + H_{PQ} \frac{1}{E - H_{QQ}} H_{QP}. \quad (24.35)$$

This is an *identity*: for any eigenvalue E of $H(a)$ in the low-energy subspace ($E \ll \pi/a'$), the Feshbach effective Hamiltonian has the same eigenvalue:

$$H(a) |\psi\rangle = E |\psi\rangle \implies H_{\text{eff}}(E) |\psi_P\rangle = E |\psi_P\rangle, \quad |\psi_P\rangle = P |\psi\rangle. \quad (24.36)$$

No approximation is involved: the Feshbach formula preserves every eigenvalue exactly. The only subtlety is that $H_{\text{eff}}(E)$ depends on E (the resolvent $1/(E - H_{QQ})$ contains E). For eigenvalues $E \ll \pi/a'$, the energy-independent approximation $H_{\text{eff}} \approx H_{PP} - H_{PQ} H_{QQ}^{-1} H_{QP}$ holds to accuracy $O(Ea')$, and the resulting effective Hamiltonian is the Wilson RG image $H(a')$ with parameters $\theta(a') = \mathcal{R}_{a \rightarrow a'}[\theta(a)]$.

Every Born-layer observable is a function of the low-energy spectrum and matrix elements of H : masses are eigenvalue differences, cross sections are S -matrix elements built from the Dyson series, decay widths are imaginary parts of self-energies. Since the Feshbach formula preserves the low-energy eigenvalues and matrix elements:

$$\boxed{\mathcal{O}[\mathcal{R}_{a \rightarrow a'}(\theta(a)), a'] = \mathcal{O}[\theta(a), a] + O(E_{\text{obs}} \cdot a'),} \quad (24.37)$$

where E_{obs} is the energy scale of the observable. The correction $O(E_{\text{obs}} \cdot a') = O(a'/\lambda)$ is the lattice artefact from the energy-independent approximation. In the exact Feshbach formula (energy-dependent), the equality is exact.

In words: the Wilson RG map produces engine parameters at the coarser lattice that give the same Born-layer outputs as the original engine at the finer lattice. Mode elimination preserves the rendering. $\square$

Proof of (B): rendering stability implies Wilson

Suppose $\theta(a)$ is a family satisfying the rendering-stability condition (L) for all $a > 0$. Fix two lattice spacings $a < a'$ and define:

$$\tilde{\theta}(a') \equiv \mathcal{R}_{a \rightarrow a'}[\theta(a)] \quad (\text{the Wilson image}). \quad (24.38)$$

By part (A):

$$\mathcal{O}[\tilde{\theta}(a'), a'] = \mathcal{O}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda). \quad (24.39)$$

By the rendering-stability hypothesis:

$$\mathcal{O}[\theta(a'), a'] = \mathcal{O}_{\text{phys}} + o(a'/\lambda). \quad (24.40)$$

Therefore:

$$\mathcal{O}[\tilde{\theta}(a'), a'] = \mathcal{O}[\theta(a'), a'] + o(a'/\lambda). \quad (24.41)$$

Both $\tilde{\theta}(a')$ and $\theta(a')$ produce the same Born-layer observables at lattice spacing a' , to accuracy $o(a'/\lambda)$.

Now invoke the *local invertibility condition*: the map $\theta \mapsto \mathcal{O}[\theta, a']$ is locally invertible at $\theta(a')$. This means: if two parameter sets produce the same observables (to some accuracy), then the parameter sets themselves agree (to the same accuracy). Formally: if $\mathcal{O}[\theta_1, a'] = \mathcal{O}[\theta_2, a'] + \epsilon$, then $\theta_1 = \theta_2 + O(\epsilon)$, provided the Jacobian $\partial\mathcal{O}/\partial\theta$ is non-degenerate.

Applying this to (24.41):

$$\boxed{\theta(a') = \mathcal{R}_{a \rightarrow a'}[\theta(a)] + o(a'/\lambda).} \quad (24.42)$$

The rendering-stable family $\theta(a)$ satisfies the Wilson RG equation, up to lattice artefacts. $\square$

The local invertibility condition

The proof of (B) requires that the map $\theta \mapsto \mathcal{O}[\theta, a']$ be locally invertible: that the Born-layer observables uniquely determine the engine parameters (locally). This is equivalent to the statement that the Jacobian matrix:

$$J_{k\alpha} = \left. \frac{\partial \mathcal{O}_k}{\partial \theta_\alpha} \right|_{\theta=\theta(a'), a=a'} \quad (24.43)$$

has full rank: the number of independent observables $\{\mathcal{O}_k\}$ is at least the number of engine parameters $\{\theta_\alpha\}$.

For a *renormalisable* theory, this is precisely the content of renormalisability: n engine parameters are fixed by n rendering conditions (Born-layer measurements), and the Jacobian is non-degenerate. For ϕ^4 : 3 parameters, 3 conditions, J is 3×3 and non-degenerate. For the Standard Model: 19 parameters, 19 conditions, J is 19×19 and non-degenerate.

For a *non-renormalisable* theory (e.g. gravity): infinitely many parameters appear at each loop order, and the number of independent rendering conditions is infinite. The Jacobian is infinite-dimensional. The equivalence (B) still holds order by order in perturbation theory (finitely many parameters at each order), but the global statement requires the full non-perturbative framework.

Remark. The local invertibility condition is the bridge between the two views. It says: the Born-layer observables carry enough information to reconstruct the engine parameters. Without it, rendering stability would be a weaker condition: many different families $\theta(a)$ could produce the same observables, and only one of them would agree with Wilson. With it, the family is unique, and it must be the Wilson family.

The equivalence, restated

	Lattice-first view	Wilson view
Defines	What $\theta(a)$ must satisfy: $\mathcal{O}[\theta(a)] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$	How $\theta(a')$ is constructed from $\theta(a)$: integrate out the shell
Asks	Does a stable $\theta(a)$ exist?	What are the fixed points and flows?
Tool	Rendering-stability condition	Feshbach / functional RG / blocking
Math	Constraint on a family of couplings	Dynamical system (RG flow)
Equivalence		
(A)	Wilson’s construction automatically satisfies rendering stability (Feshbach preserves eigenvalues).	
(B)	Rendering stability, plus local invertibility, implies the Wilson flow (up to $o(a'/\lambda)$).	

An analogy. Consider a sequence $\{x_n\}$ defined by the requirement $|f(x_n) - c| < \epsilon_n$ (the “output-stability” view) versus a sequence defined by the recursion $x_{n+1} = g(x_n)$ (the “dynamical” view). If g is the Newton-Raphson iteration for the equation $f(x) = c$, then the recursion generates the unique sequence converging to the root of $f(x) = c$ —and any sequence with $f(x_n) \rightarrow c$ must asymptotically agree with the Newton-Raphson iterates, provided f' is non-degenerate. The output-stability view defines what the sequence must achieve; the dynamical view constructs the sequence; the equivalence says they produce the same result.

The payoff for the ML reader. The lattice-first view is the natural one for machine learning: it says “tune the model parameters so the test-set performance is resolution-invariant.” The Wilson view is the natural one for physics: it says “integrate out the fine-grained degrees of freedom to obtain an effective coarse-grained model.” The equivalence theorem says these are two descriptions of the same mathematical object: the RG flow is the unique parameter trajectory that maintains performance across resolutions.

In neural-network language: Wilson’s RG is like progressive distillation—training a smaller model to reproduce the outputs of a larger one. Rendering stability is like the requirement that the distilled model has the same test-set accuracy. The equivalence says: distillation is the unique procedure that maintains accuracy, and any accuracy-preserving reparameterisation must agree with distillation.

24.7 Emergence of operators: zero couplings that become nonzero

The equivalence theorem (Section 24.6) established that the Wilson RG flow generates the unique family $\theta(a)$ satisfying rendering stability. But $\theta(a)$ may contain *more couplings* than the original theory at spacing a . The Feshbach shell integral, or equivalently the rendering-stability condition at a coarser lattice, can generate operators that were absent in the original engine. Understanding when this happens—and when it matters—is essential for both views.

The mechanism: Wilson generates all operators compatible with the symmetries

Start with the engine $H(a)$ containing two operators: a free kinetic term and the ϕ^4 interaction:

$$H(a) = \sum_{\mathbf{k}} \omega_{\mathbf{k}} a_{\mathbf{k}}^{\dagger} a_{\mathbf{k}} + \frac{\lambda(a)}{4!} \int d^3x : \hat{\phi}^4 : . \quad (24.44)$$

No ϕ^6 , no $(\partial\phi)^4$, no higher-order terms. Their couplings are exactly zero at scale a : $c_6(a) = c_8(a) = \dots = 0$.

Now integrate out the momentum shell $\Lambda' < |\mathbf{k}| < \Lambda$ via the Feshbach formula. The effective Hamiltonian at the coarser lattice a' :

$$H_{\text{eff}} = H_{PP} + H_{PQ} \frac{1}{E - H_{QQ}} H_{QP} + \dots \quad (24.45)$$

The second-order term $H_{PQ}(E - H_{QQ})^{-1}H_{QP}$ connects low-momentum states to high-momentum intermediate states and back. The resulting operator, expanded in local field operators, contains *every* monomial of $\hat{\phi}$ compatible with the symmetries:

$$H(a') = \sum_{\mathbf{k}} \omega'_{\mathbf{k}} a_{\mathbf{k}}^{\dagger} a_{\mathbf{k}} + \frac{\lambda(a')}{4!} \int : \phi^4 : + \frac{c_6(a')}{6!} \int : \phi^6 : + \frac{c_8(a')}{8!} \int : \phi^8 : + \dots \quad (24.46)$$

The couplings $c_6(a')$, $c_8(a')$, $\dots$ are nonzero, even though they were zero at scale a . The shell integration has **induced new operators**.

The explicit one-loop values for ϕ^4 theory:

$$c_6(a') = \frac{15 \lambda(a)^3}{(4\pi)^2} a'^2 + O(\lambda^4), \quad (24.47)$$

$$c_8(a') = O(\lambda^4 a'^4). \quad (24.48)$$

(The factor of 15 counts the Wick contractions of three ϕ^4 vertices that produce a ϕ^6 topology; the factor a'^2 is dimensional analysis: c_6 has dimension $[\text{length}]^2$ in $d = 4$.)

This is the Wilson-side statement: the RG flow generates all operators compatible with the symmetries, regardless of their initial values.

The rendering-stability side: which induced operators matter?

Now examine the same phenomenon from the rendering-stability view. The question: if we start with $\theta(a) = (m^2(a), \lambda(a), Z(a))$ (three parameters, no ϕ^6), and change to a' , do we need $c_6(a')$ to maintain rendering stability?

The answer depends on the operator's **scaling dimension** relative to the spacetime dimension d .

Irrelevant operators (dimension $> d$): generated but invisible. The operator ϕ^6 in $d = 4$ has mass dimension $6 > 4$. Its effect on a Born-layer observable at observation scale λ is:

$$\delta\mathcal{O}|_{\phi^6} = c_6(a') \cdot \lambda^{-2} \cdot (\text{matrix element}) \sim \frac{\lambda(a)^3}{(4\pi)^2} \frac{a'^2}{\lambda^2} = O\left(\frac{a'}{\lambda}\right)^2. \quad (24.49)$$

This is $o(a'/\lambda)$: it vanishes in the rendering-stability limit. The ϕ^6 operator is generated by the RG, but its Born-layer effect is a lattice artefact, indistinguishable from the discretisation error.

From the rendering-stability view: the induced ϕ^6 coupling does not need to be included in $\theta(a')$. It can be absorbed into the $o(a/\lambda)$ correction. Three engine parameters suffice.

Marginal operators (dimension $= d$): generated and visible. The operator ϕ^4 in $d = 4$ has mass dimension $4 = d$. Its effect on a Born-layer observable is:

$$\delta\mathcal{O}|_{\phi^4} = \delta\lambda \cdot (\text{matrix element}) \sim \frac{\lambda(a)^2}{(4\pi)^2} \ln \frac{a'}{a}. \quad (24.50)$$

This is $O(\ln(a'/a))$: it does NOT vanish in the rendering-stability limit. The induced shift $\delta\lambda$ must be absorbed into $\lambda(a')$ to maintain rendering stability.

Relevant operators (dimension $< d$): generated and dominant. The operator $m^2\phi^2$ has mass dimension $2 < 4$. Its induced shift is:

$$\delta m^2 = \frac{\lambda(a)}{(4\pi)^2} \frac{c}{a'^2} \left(1 - \frac{a'^2}{a^2}\right) = O(1/a'^2). \quad (24.51)$$

This grows as $1/a'^2$: it dominates over all other corrections. It must be absorbed into $m^2(a')$.

Summary: the RG generates all operators; rendering stability filters by relevance.

Operator	Dim.	Induced pling	cou-	Born effect	Rendering stability
ϕ^2 (mass)	2	$\delta m^2 \sim 1/a'^2$		$O(1/\lambda^2)$: dominant	Must include in $\theta(a')$
ϕ^4 (quartic)	4	$\delta\lambda \sim \ln(a'/a)$		$O(\ln)$: visible	Must include in $\theta(a')$
$(\partial\phi)^2$ (wave fn.)	4	$\delta Z \sim \ln(a'/a)$		$O(\ln)$: visible	Must include in $\theta(a')$
ϕ^6	6	$c_6 \sim \lambda^3 a'^2$		$O(a'^2/\lambda^2)$: invisible	Absorbed in $o(a'/\lambda)$
ϕ^8	8	$c_8 \sim \lambda^4 a'^4$		$O(a'^4/\lambda^4)$: invisible	Absorbed in $o(a'/\lambda)$

Wilson generates everything. Rendering stability keeps only what matters: the relevant and marginal operators. The irrelevant operators are present in the Fock layer but invisible in the

Born layer — they are lattice artefacts, suppressed by powers of a/λ . This is precisely the content of *universality*: the long-distance physics is insensitive to the short-distance details (the irrelevant operators), and the only parameters that matter are the relevant and marginal couplings.

Three scenarios for emergent operators

The interplay between Wilson generation and rendering-stability filtering produces three qualitatively different scenarios.

Scenario 1: generated and dies. An irrelevant operator ($\phi^6, \phi^8, (\partial\phi)^4, \dots$) is generated by the RG with a nonzero coupling $c_n(a') \propto a'^{n-d}$. Its Born-layer effect scales as $(a'/\lambda)^{n-d} \rightarrow 0$. The operator emerges in the Fock layer and dies in the Born layer. It exists at every finite a' but contributes only a lattice artefact to the observables.

This is the generic case: most of the operators generated by the RG are irrelevant and invisible. They populate the effective Hamiltonian $H(a')$ but do not affect the physics.

In ML terms: these are high-frequency features in the hidden representation that do not affect the model's predictions at the output layer. They are present in the weights but absent from the test-set performance.

Scenario 2: generated and survives. A relevant or marginal operator is absent at scale a (perhaps forbidden by a symmetry), but the symmetry is broken at scale a' by the lattice discretisation or by another interaction. The operator is generated with an unsuppressed coupling, and its Born-layer effect is $O(1)$ or $O(\ln)$.

Example: fermion mass from broken chiral symmetry. Chiral symmetry forbids the fermion mass term $m\bar{\psi}\psi$. If the lattice preserves chiral symmetry (overlap or domain-wall fermions): the mass remains zero, $m(a') = 0$ for all a' . If the lattice breaks chiral symmetry (Wilson fermions with the $r\bar{\psi}\partial^2\psi$ term): a mass $\delta m \sim r/a'$ is generated, and it is relevant (dimension $3 < 4$). The fermion acquires a mass of order $1/a'$ — a lattice artefact that *does not die*, because the operator is relevant. It must be cancelled by tuning $m(a')$ to restore the physical mass.

This is why the choice of lattice discretisation matters for fermions (Section 24.11): it determines which symmetries are preserved, and hence which operators can be generated. The rendering-stability condition requires the same number of engine parameters as the number of relevant and marginal operators that the lattice generates — which depends on the symmetries of the lattice action.

Example: cosmological constant from matter loops. The unit operator $\mathbf{1}$ has dimension $0 < 4$: it is relevant. Even if the vacuum energy is zero at scale a ($E_0(a) = 0$), the shell integration generates $\delta E_0 \sim N_{\text{modes}} \cdot \omega_{\text{shell}}/2 \sim 1/a'^4$. This is the a_0 Seeley–DeWitt coefficient (Section 23.10). The cosmological constant is a relevant operator that is always generated; it must always be included in $\theta(a')$.

Scenario 3: generated and dominates (emergence from nothing). In special cases, an operator that is entirely absent from the original action can become the dominant term at long distances. The operator is not part of the UV engine; it is an emergent feature of the IR rendering.

Example: induced gravity. Consider a matter field theory on a curved background $g_{\mu\nu}$, with no gravitational kinetic term in the action (no $\int \sqrt{g} R$). The tree-level action has $G_N^{-1} = 0$: gravity is absent. But the a_1 Seeley–DeWitt coefficient (Section 23.10) generates:

$$\Gamma_{\text{ind}} = \frac{N_{\text{dof}}}{192\pi^2 a^2} \int d^4x \sqrt{g} R + \dots \quad (24.52)$$

This is a relevant operator (dimension 2 on a 4-manifold): it has a coefficient $\propto 1/a^2$ that grows as the lattice is refined. The induced Newton constant is $G_N^{-1} \propto N_{\text{dof}}/a^2$ — nonzero, large, and entirely generated by the RG.

Gravity has emerged from a theory that had no gravity. The gravitational kinetic term was not put in by hand; it was induced by integrating out matter fluctuations. This is Sakharov’s induced gravity (1967): the metric $g_{\mu\nu}$ is an interface observable (it determines distances and angles on the Born-layer screen), and its dynamics (general relativity) emerges from the matter engine’s one-loop effective action.

At two loops: the a_2 coefficient generates the curvature-squared terms R^2 and $C_{\mu\nu\rho\sigma}^2$, which are marginal operators. They contribute to the running of the gravitational couplings at short distances. At higher loops: still higher-order curvature invariants are generated.

From the rendering-stability view: the question is whether the induced gravitational action, with its tower of curvature invariants, is rendering-stable with finitely many parameters. If the induced gravitational RG flow reaches a UV fixed point (asymptotic safety): finitely many parameters suffice. If not: the effective description breaks down at some scale, and gravity is a purely emergent, long-distance phenomenon with no UV completion.

The equivalence theorem, extended

The equivalence theorem of Section 24.6 extends to induced operators.

Wilson side: the Feshbach formula generates all operators compatible with the symmetries. The induced couplings are determined by the shell integral.

Rendering-stability side: the rendering-stability condition at scale a' requires parameters for all relevant and marginal operators, whether or not they were present at scale a . Their values are determined by the condition $\mathcal{O}[\theta(a'), a'] = \mathcal{O}_{\text{phys}} + o(a'/\lambda)$.

Equivalence: the Wilson-generated couplings are the unique values satisfying rendering stability (by the local invertibility argument). The irrelevant operators generated by Wilson are exactly those absorbed into the $o(a'/\lambda)$ correction of the rendering-stability condition.

$$\boxed{\begin{array}{ccc} \underbrace{\text{Wilson: generate all operators}}_{\text{the Feshbach formula}} & \longleftrightarrow & \underbrace{\text{Rendering stability: keep relevant + marginal, discard irrelevant as } o(a/\lambda)}_{\text{the rendering-stability condition}} \end{array}} \quad (24.53)$$

Universality is the statement that the two sides agree: the long-distance physics depends only on the relevant and marginal couplings (which the rendering-stability condition requires), and the irrelevant couplings (which Wilson generates but the Born layer cannot see) are lattice artefacts.

In ML terms. Training a large model and distilling it into a smaller one: the large model (fine lattice) has many parameters, including high-frequency features (irrelevant operators). The distilled model (coarse lattice) retains only the features that affect the predictions (relevant and marginal). The distillation loss is $o(\text{resolution ratio})$: the high-frequency features wash out. Wilson’s RG is the distillation; rendering stability is the requirement that the distilled model’s predictions match the original’s. Operator emergence is the observation that the distilled model may use different features (operators) than the original — features that emerge from the distillation process itself, not from the original parameterisation.

The $\langle JJ \rangle$ tower for gauge fields. The conserved current $J^\mu = i(\phi^* \partial^\mu \phi - \phi \partial^\mu \phi^*)$ has mass dimension $[J^\mu] = d - 1 = 3$ in $d = 4$. The current couples to the gauge field via $\int A_\mu J^\mu$, and the one-loop effective action for A_μ is built from the current-current correlator:

$$\Pi^{\mu\nu}(p) = \langle J^\mu(p) J^\nu(-p) \rangle_{\text{conn}}. \quad (24.54)$$

Current conservation $\partial_\mu J^\mu = 0$ forces the Ward identity $p_\mu \Pi^{\mu\nu} = 0$, which constrains the tensor structure:

$$\Pi^{\mu\nu}(p) = (p^2 \eta^{\mu\nu} - p^\mu p^\nu) \Pi(p^2). \quad (24.55)$$

The overall transverse projector $(p^2 \eta^{\mu\nu} - p^\mu p^\nu)$ vanishes at $p = 0$. Expand $\Pi(p^2)$ in powers of p^2 :

Order	Gauge operator	Coupling dim.	Class	Status
p^0	$A_\mu A^\mu$ (mass term)	$[m_A^2] = 2$	Relevant	Killed by Ward identity
p^2	$F_{\mu\nu} F^{\mu\nu}$ (kinetic)	$[1/g^2] = 0$	Marginal	Induced; $1/g^2 \sim \ln(1/a)$
p^4	$(D_\mu F^{\mu\nu})^2$, etc.	$[c] = -2$	Irrelevant	$c \sim a^2$; dies
p^6	Higher-derivative F terms	$[c'] = -4$	Irrelevant	$c' \sim a^4$; dies

The crucial row is p^0 : the gauge boson mass $m_A^2 A_\mu A^\mu$ would be a relevant operator (dimension $2 < d$), generating a power-law a -dependence $m_A^2 \propto 1/a^2$. But the Ward identity kills it: the transverse structure (24.55) forces $\Pi^{\mu\nu}(0) = 0$. No mass is generated. The condition number is $\kappa = 0$ (Section ??): the critical surface for $m_A \neq 0$ does not exist within the gauge-invariant theory.

The first surviving term is p^2 : the kinetic term $F_{\mu\nu} F^{\mu\nu}$, with dimensionless coupling $1/g^2$. This is marginal. Everything beyond ($p^4, p^6, \dots$) is irrelevant. Therefore:

The $\langle JJ \rangle$ correlator generates exactly one marginal operator and nothing relevant.

The structural comparison. Place the $\langle JJ \rangle$ and $\langle TT \rangle$ towers side by side:

Order	Gauge ($\langle JJ \rangle$)	Gravity ($\langle TT \rangle$)
p^0	$m_A^2 A^2$: relevant, killed by Ward	Λ : relevant, not killed ($\kappa \sim 10^{120}$)
p^2	$F_{\mu\nu} F^{\mu\nu}$: marginal (the kinetic term)	$M_P^2 R$: relevant ($\kappa \gg 1$ for Planck hierarchy)
p^4	$(D_\mu F^{\mu\nu})^2$: irrelevant	R^2, C^2 : marginal
p^6	Irrelevant	R^3 : irrelevant

The pattern is now clear:

For gauge theory: the Ward identity kills the only relevant operator (the mass term). The kinetic term F^2 is marginal. Everything else is irrelevant. Result: **one marginal coupling** (g), logarithmic running, renormalisable with finitely many parameters.

For gravity: no Ward identity kills the cosmological constant or the Einstein–Hilbert term. Both are relevant. The curvature-squared terms are marginal. Result: **two relevant couplings** (Λ, M_P^2) with power-law running, **two marginal couplings** (c_{R^2}, c_{C^2}) with logarithmic running, and irrelevant operators starting at p^6 . Whether this tower closes (finitely many parameters suffice for rendering stability, i.e., asymptotic safety) is an open question.

The root cause: the dimension of the background field. Why does the tower shift? Because the background field has different mass dimension:

$$[A_\mu] = 1, \quad [g_{\mu\nu}] = 0. \quad (24.56)$$

The kinetic term is built from two derivatives of the background field:

$$[F_{\mu\nu}] = [\partial A] = [A] + 1 = 2, \quad [F^2] = 4 = d \Rightarrow \text{marginal}. \quad (24.57)$$

$$[R] = [\partial^2 g] = [g] + 2 = 2, \quad [R] = 2 < d \Rightarrow \text{relevant}. \quad (24.58)$$

For gauge fields: two derivatives of a dimension-1 field give a dimension-4 object. Squaring it gives dimension 8: irrelevant. The kinetic term F^2 is the unique marginal operator.

For gravity: two derivatives of a dimension-0 field give a dimension-2 object. Squaring it gives dimension 4: marginal. The kinetic term R is relevant, and the first correction R^2 reaches marginal. The “promotion” of the first correction from irrelevant (gauge) to marginal (gravity) is the structural reason gravity generates more counterterms.

In one formula: the number of marginal gravitational operators exceeds the number of marginal gauge operators because the metric’s dimension is lower than the gauge field’s dimension by exactly 1, which shifts the entire tower down by 2 units (since the operators are quadratic in the field strength/curvature).

The Fock-Born summary. Both gauge dynamics and gravity can emerge from a scalar engine through the mechanism: composite current/stress tensor $\rightarrow$ correlator $\rightarrow$ induced kinetic term. The emergent dynamics is classified by the tower of induced operators:

	[source]	Relevant ops.	Marginal ops.	Renormalisable?
Gauge (A_μ)	1	0 (Ward kills mass)	1 (F^2)	Yes: 1 coupling
Gravity ($g_{\mu\nu}$)	0	2 (Λ , R)	2 (R^2 , C^2)	Open (asymptotic safety?)

The scalar engine generates both gauge forces and gravity through the same mechanism. The difference in their renormalisability is not a difference in the mechanism but in the dimension of the background field: $[A_\mu] = 1$ gives a renormalisable theory; $[g_{\mu\nu}] = 0$ gives a theory whose renormalisability depends on the existence of a UV fixed point. In the lattice-first view, both theories are well-defined at finite a regardless; the question of renormalisability is about the continuum extrapolation $a \rightarrow 0$.

The lattice-first verdict. From the lattice-first view, the structural difference between gauge theory and gravity is real but less dramatic than the continuum narrative suggests.

On the lattice at spacing a , every induced coupling is a finite number:

$$\Lambda(a) \sim \frac{c_0}{a^4}, \quad M_P^2(a) \sim \frac{c_1}{a^2}, \quad c_{R^2}(a) \sim c_2 \ln \frac{1}{ma}, \quad c_{R^3}(a) \sim c_3 a^2. \quad (24.59)$$

All finite. No divergence. The tower of gravitational operators, from relevant through marginal through irrelevant, is a finite list at every finite a (the lattice has finitely many modes, so only finitely many independent operators exist).

The rendering-stability condition at finite a :

$$\mathcal{O}_{\text{Born}}[\Lambda(a), M_P^2(a), c_{R^2}(a), \dots; a] = \mathcal{O}_{\text{phys}} + o(a/\lambda), \quad (24.60)$$

is a finite system of equations for a finite number of parameters. It is satisfied at every finite a . The engine runs.

Asymptotic safety asks: does the family $\theta(a)$ converge to a fixed point as $a \rightarrow 0$? If yes, the continuum rendering exists with finitely many free parameters. If no, the continuum rendering does not exist—but the physics at finite a is unchanged.

In the classification of Section ??: asymptotic safety is **type B**—a question about the rendering’s infinite-resolution limit. It does not affect the engine. The engine has a finite lattice, finite parameters, finite observables, and finite rendering stability at every finite a . Whether the limit $a \rightarrow 0$ exists is a mathematical question, not a physical one.

The continuum narrative says: “gravity is non-renormalisable, so we need asymptotic safety (or string theory, or something) to make sense of it at short distances.” The lattice-first narrative says: “gravity is well-defined on the lattice at every a . The ‘non-renormalisability’ is the statement that the continuum extrapolation requires infinitely many parameters. But we never take $a \rightarrow 0$. The engine operates at finite a —perhaps $a \sim \ell_P$ —and the question of what happens at $a = 0$ is not a physics question.”

What IS physically meaningful. The lattice-first view does not dismiss everything about the gauge-gravity comparison. The following are genuine physical statements, independent of the continuum extrapolation:

Gauge boson masses are protected ($\kappa = 0$). The Ward identity $p_\mu \Pi^{\mu\nu} = 0$ holds on the lattice (it is a consequence of exact gauge invariance of the Wilson plaquette action). The photon mass is exactly zero at every a . No tuning is needed.

The cosmological constant is not protected ($\kappa \sim 10^{120}$). No lattice Ward identity kills the p^0 term of $\langle TT \rangle$. The vacuum energy $\Lambda(a) \sim 1/a^4$ is generated at every a , and maintaining a small physical Λ_{obs} requires $\kappa \sim 10^{120}$ — a genuine numerical instability of the engine, independent of perturbation theory and independent of the continuum limit.

The R^2 couplings run logarithmically. The marginal gravitational couplings $c_{R^2}(a)$ and $c_{C^2}(a)$ are dimensionless and run with $\ln(1/a)$, just like gauge couplings. Their values are measurable in principle (they affect gravitational scattering at energies $E \sim M_P$), though in practice the effects are $O(E^2/M_P^2)$ and far below current experimental sensitivity.

These are statements about the engine at finite a . They do not require the continuum limit. They do not require asymptotic safety. They are the physical content of the $\langle JJ \rangle$ vs. $\langle TT \rangle$ comparison, stated in the correct language.

24.8 Weinberg’s effective field theory programme

Weinberg (1979) elevated a practical technique—organising interactions by mass dimension—into a philosophy of physics. The result is a programme that is powerful, general, and (in the lattice-first view) incomplete in a specific, identifiable way. We develop the programme technically and give four applications before connecting it to the lattice-first framework.

The theorem

Weinberg’s folk theorem. *At energies $E \ll \Lambda$, the most general Lagrangian consistent with the symmetries G and containing the light degrees of freedom produces the most general S -matrix consistent with analyticity, unitarity, cluster decomposition, and G .*

This is not a definition but a theorem (proved under varying assumptions by Weinberg, 1979; Leutwyler, 1994; Polchinski, 1984 via the Wilsonian RG). The content: you cannot miss any

physical effect at low energy by writing down all operators. If an effect exists and is consistent with the symmetries, it will appear as a term in the EFT Lagrangian.

The converse also holds: every term in the Lagrangian corresponds to a physical effect. There are no “redundant” operators (after using the equations of motion to reduce the operator basis).

The procedure

Step 1: identify the degrees of freedom. List the particles with mass $m \ll \Lambda$ (the “light” or “active” degrees of freedom). Particles with $m \gtrsim \Lambda$ are “heavy” or “integrated out”: their effects appear as local operators involving only the light fields.

Step 2: identify the symmetries. List all symmetries of the low-energy theory: Lorentz invariance, gauge symmetries, global symmetries (exact and approximate), discrete symmetries (C , P , T).

Step 3: write all operators, organised by dimension. Construct every Lorentz-scalar, G -invariant local operator $\mathcal{O}_i$ from the light fields and their derivatives. Organise by mass dimension Δ_i . The EFT Lagrangian:

$$\mathcal{L}_{\text{EFT}} = \sum_{\Delta_i \leq d} c_i \mathcal{O}_i + \sum_{\Delta_i = d+1} \frac{c_j}{\Lambda} \mathcal{O}_j + \sum_{\Delta_i = d+2} \frac{c_k}{\Lambda^2} \mathcal{O}_k + \cdots \quad (24.61)$$

The coefficients c_i are dimensionless (for $\Delta_i \leq d$) or carry powers of $1/\Lambda$ (for $\Delta_i > d$). They are *free parameters*, to be determined by matching to experiment or to a UV theory.

Step 4: power counting. At energy $E \ll \Lambda$, the contribution of an operator $\mathcal{O}_i$ with $\Delta_i = d + n$ is suppressed by $(E/\Lambda)^n$ relative to the leading terms. The EFT is a systematic expansion in E/Λ :

Leading order (LO): operators with $\Delta_i \leq d$ (relevant and marginal).

Next-to-leading order (NLO): operators with $\Delta_i = d + 1$, suppressed by E/Λ .

Next-to-next-to-leading (NNLO): $\Delta_i = d + 2$, suppressed by $(E/\Lambda)^2$.

At each order: a finite number of operators and a finite number of free coefficients. Predictions at order $(E/\Lambda)^n$ require determining the coefficients at all orders $\leq n$.

Step 5: matching. If the UV theory is known (e.g., the Standard Model for the Fermi theory): compute the coefficients c_i by requiring the EFT amplitudes to agree with the UV amplitudes at energies $E \sim \Lambda$ (“matching”). If the UV theory is unknown: the coefficients are free parameters, determined by experiment.

The power-counting formula

For a Feynman diagram in the EFT with L loops, I internal lines, V_i vertices of type i (each from an operator of dimension Δ_i), the amplitude scales as:

$$\mathcal{A} \sim E^D, \quad D = d + \sum_i V_i(\Delta_i - d) + L(d - 2) - 2I + (\text{external}). \quad (24.62)$$

Using the topological identity $L = I - \sum_i V_i + 1$:

$$D = d + (d - 2)L + \sum_i V_i(\Delta_i - 2). \quad (24.63)$$

(In $d = 4$: $D = 4 + 2L + \sum_i V_i(\Delta_i - 2)$.)

Each additional loop adds $d - 2 = 2$ powers of E . Each irrelevant vertex ($\Delta_i > d$) adds $\Delta_i - d$ powers. The leading contribution (D minimal) comes from tree diagrams ($L = 0$) with only renormalisable vertices ($\Delta_i \leq d$).

Application 1: the Fermi theory of weak interactions

UV theory: the Standard Model with $W^\pm$ and Z^0 bosons ($m_W \approx 80 \text{ GeV}$).

EFT scale: $\Lambda = m_W$. At $E \ll m_W$: the W and Z are “integrated out.”

Light degrees of freedom: quarks (u, d, s), leptons (e, μ, ν), photon, gluons.

Symmetries: $SU(3)_c \times U(1)_{\text{em}}$, Lorentz.

Leading operator: the four-fermion interaction, dimension 6:

$$\mathcal{L}_{\text{Fermi}} = -\frac{4G_F}{\sqrt{2}} (\bar{\psi}_1 \gamma^\mu P_L \psi_2) (\bar{\psi}_3 \gamma_\mu P_L \psi_4) + \text{h.c.}, \quad (24.64)$$

where $G_F = g^2/(4\sqrt{2}m_W^2) \approx 1.166 \times 10^{-5} \text{ GeV}^{-2}$ is the Fermi constant and $P_L = \frac{1}{2}(1 - \gamma_5)$.

Matching: at tree level, the W -exchange amplitude is:

$$\mathcal{A}_W = \frac{g^2}{q^2 - m_W^2} \xrightarrow{q^2 \ll m_W^2} -\frac{g^2}{m_W^2} \left(1 + \frac{q^2}{m_W^2} + \dots\right). \quad (24.65)$$

The leading term $-g^2/m_W^2$ gives G_F . The correction q^2/m_W^2 is an NLO effect, suppressed by $(E/m_W)^2$.

Power counting: the Fermi vertex has $\Delta = 6 = d + 2$, so each insertion contributes $(E/m_W)^2$. A one-loop diagram with one Fermi vertex: $D = 4 + 2 + 2 = 8$, i.e., $\mathcal{A} \sim E^8/m_W^4$. Corrections are systematically suppressed.

Prediction: muon decay width $\Gamma(\mu \rightarrow e \bar{\nu}_e \nu_\mu) = G_F^2 m_\mu^5/(192\pi^3)$, accurate to $O(m_\mu^2/m_W^2) \sim 10^{-6}$.

Application 2: chiral perturbation theory

UV theory: QCD with $N_f = 2$ or 3 light quark flavours.

EFT scale: $\Lambda \sim \Lambda_{\text{QCD}} \approx 1 \text{ GeV}$ (or $4\pi f_\pi \approx 1.2 \text{ GeV}$).

Light degrees of freedom: pions π^a (pseudo-Nambu–Goldstone bosons of the spontaneously broken chiral symmetry $SU(N_f)_L \times SU(N_f)_R \rightarrow SU(N_f)_V$). For $N_f = 3$: also K and η .

Symmetries: the nonlinearly realised chiral symmetry. The pions are parameterised by $U(x) = \exp(2i\pi^a(x)T^a/f_\pi) \in SU(N_f)$, transforming as $U \rightarrow LUR^\dagger$ under $(L, R) \in SU(N_f)_L \times SU(N_f)_R$.

Operator counting: the chiral Lagrangian is organised by powers of ∂/Λ (or equivalently p/Λ , since pions are derivatively coupled):

LO (p^2):

$$\mathcal{L}_2 = \frac{f_\pi^2}{4} \text{tr}(\partial_\mu U^\dagger \partial^\mu U) + \frac{f_\pi^2 B_0}{2} \text{tr}(M_q^\dagger U + U^\dagger M_q), \quad (24.66)$$

where $f_\pi \approx 93 \text{ MeV}$ is the pion decay constant, B_0 is related to the quark condensate, and $M_q = \text{diag}(m_u, m_d)$ is the quark mass matrix. The second term gives the pion mass: $m_\pi^2 = B_0(m_u + m_d)$ (the Gell-Mann–Oakes–Renner relation).

NLO (p^4): 10 independent operators for $N_f = 3$ (the Gasser–Leutwyler Lagrangian $\mathcal{L}_4$), with 10 free low-energy constants $L_1, \dots, L_{10}$. These encode effects of heavier resonances ($\rho, \omega, a_1, \dots$) that have been integrated out.

NNLO (p^6): 90 independent operators for $N_f = 3$, with 90 free constants.

Predictions at LO: $\pi\pi$ scattering lengths $a_0^0 = 7m_\pi^2/(32\pi f_\pi^2) \approx 0.16$ (Weinberg, 1966); confirmed experimentally to $\sim 5\%$. NLO corrections (one-loop + L_i) improve the prediction to $\sim 1\%$.

The counting is non-standard: pion interactions are derivatively coupled ($\mathcal{L} \sim (\partial\pi)^2/f_\pi^2$), so the expansion parameter is $p^2/(4\pi f_\pi)^2$, not p/Λ . Each loop adds $p^2/(4\pi f_\pi)^2$, not p^2 .

Application 3: gravity as EFT

UV theory: unknown.

EFT scale: $\Lambda = M_P \approx 2.4 \times 10^{18}$ GeV (the reduced Planck mass).

Light degrees of freedom: the metric $g_{\mu\nu}$ (the massless spin-2 graviton).

Symmetries: $\text{Diff}(M)$.

Operator expansion:

$$\mathcal{L}_{\text{grav}} = \sqrt{g} \left[-\Lambda_{\text{cc}} + \frac{M_P^2}{2} R + c_1 R^2 + c_2 R_{\mu\nu} R^{\mu\nu} + \frac{c_3}{M_P^2} R^3 + \dots \right]. \quad (24.67)$$

Λ_{cc} (dimension 4): the cosmological constant. Two engine parameters: Λ_{cc} and M_P^2 .

R (dimension 2 in the operator, but the coupling M_P^2 has dimension 2): relevant.

R^2 , $R_{\mu\nu}^2$ (dimension 4): marginal. Two new free coefficients c_1 , c_2 . (The Gauss–Bonnet combination is topological in $d = 4$.)

R^3 (dimension 6): irrelevant, suppressed by $1/M_P^2$.

Predictions (Donoghue, 1994): quantum corrections to the Newtonian potential:

$$V(r) = -\frac{Gm_1 m_2}{r} \left[1 + \frac{3G(m_1 + m_2)}{rc^2} + \frac{41}{10\pi} \frac{G\hbar}{r^2 c^3} + \dots \right]. \quad (24.68)$$

The first correction ($\propto G/r$) is the classical post-Newtonian term (from the nonlinearity of GR). The second correction ($\propto G\hbar/r^2$) is a genuine quantum effect: a one-loop graviton contribution, computed within the EFT. It is parameter-free at this order (no unknown coefficients from c_1 , c_2 —they contribute only at shorter distances).

This is the power of Weinberg’s programme applied to gravity: even though the UV completion is unknown, the low-energy quantum corrections are calculable and unambiguous.

Breakdown: the EFT expansion fails at $r \sim \ell_P = 1/M_P$ (the Planck length), where the corrections $(G\hbar/r^2)^n$ are no longer suppressed.

Application 4: the Standard Model as an EFT (SMEFT)

UV theory: unknown (whatever lies beyond the SM).

EFT scale: $\Lambda = \Lambda_{\text{NP}}$ (the new-physics scale, currently $\Lambda_{\text{NP}} \gtrsim \text{few TeV}$ from LHC bounds).

Light degrees of freedom: the SM particles (quarks, leptons, gauge bosons, Higgs).

Symmetries: $SU(3)_c \times SU(2)_L \times U(1)_Y$ gauge symmetry.

Operator expansion: the SM Lagrangian is the LO ($\Delta \leq 4$); the corrections are:

Dimension 5 ($\Delta = 5$): one operator (the Weinberg operator $\frac{c_5}{\Lambda}(LH)(LH)$, which generates Majorana neutrino masses after electroweak symmetry breaking).

Dimension 6 ($\Delta = 6$): 2499 independent operators in the “Warsaw basis” (Grzadkowski et al., 2010), parameterising all possible deviations from the SM at order $1/\Lambda^2$. These include four-fermion operators, anomalous gauge-boson couplings, modified Higgs couplings, dipole operators, etc.

Dimension 8: $\sim 44,000$ operators (counted by Henning et al., 2017).

The SMEFT programme at the LHC: measure the Wilson coefficients of the dimension-6 operators from deviations in Higgs production rates, diboson cross sections, top-quark observables, and electroweak precision data. Current bounds: $|c_i/\Lambda^2| \lesssim (1 \text{ TeV})^{-2}$ for most operators. No significant deviation from the SM has been observed.

Weinberg’s programme: strengths and the lattice-first critique

The strengths are genuine. Weinberg’s programme is the most general framework for low-energy physics. It makes predictions with minimal assumptions (only the symmetries and the light degrees of freedom). It has been spectacularly successful: chiral perturbation theory predicts pion scattering lengths to 1%; the Fermi theory predicts muon decay to 10^{-6} ; the gravitational EFT predicts quantum corrections to Newton’s law that are unambiguous and parameter-free at leading order.

The limitations are specific. From the lattice-first view, Weinberg’s programme has three blind spots (developed in Section ??):

1. *The coefficients are free.* Weinberg’s programme parameterises ignorance: the c_i encode the effects of the UV physics, but the programme cannot compute them. In the lattice-first view, the coefficients are computed from the engine parameters $\theta(a)$ via the Seeley–DeWitt expansion. The engine is more predictive: it determines the c_i , and it predicts relations among them (e.g., c_1/c_2 for the gravitational R^2 operators is a definite function of the matter content).
2. *The symmetries are assumed, not derived.* Weinberg assumes G ($\text{Diff}(M)$ for gravity, $SU(3) \times SU(2) \times U(1)$ for the SM). The lattice-first view derives G from the tensor structure of the composite operators (Section ??): a rank-1 composite begets gauge invariance; a rank-2 symmetric composite begets $\text{Diff}(M)$. The symmetry is a consequence of the engine, not an axiom.
3. *The programme is blind to the polishing.* The EFT operators are G -invariant by construction. The programme cannot see the G -breaking corrections $\Delta\Gamma$ that the exact theory (the HST on the lattice) contains (Section ??). In weak fields: these corrections are $o(a/\lambda)$ and invisible. In strong fields: they are $O(1)$ and the EFT breaks down. Weinberg’s programme cannot predict where the breakdown occurs or what replaces the EFT at short distances—because it is, by design, agnostic about the engine.

The honest summary. Weinberg’s programme is the correct description of the rendering at $\lambda \gg a$. Its limitations are not errors but design choices: it trades knowledge of the engine for generality. When the engine is unknown (gravity, BSM physics): the programme is the best available tool. When the engine is known (the lattice, the SM at energies below m_W): the programme is correct but less predictive than a direct computation from the engine.

The lattice-first view does not replace Weinberg’s programme; it provides the engine that Weinberg’s programme parameterises.

24.9 Weinberg’s E/Λ vs lattice-first a/λ : a term-by-term comparison

The two frameworks organise the same physics in parallel expansions. The dictionary is:

Weinberg	Lattice-first
Cutoff Λ	Lattice momentum cutoff π/a
Observation energy E	Inverse observation scale $1/\lambda$ (the “pixel size”)
Expansion parameter E/Λ	Expansion parameter a/λ
Light degrees of freedom	Low-momentum modes ($ k < 1/\lambda$)
Heavy (integrated out)	High-momentum modes ($1/\lambda < k < \pi/a$)
EFT Lagrangian $\sum c_i \mathcal{O}_i$	Effective action $\Gamma[\theta(a), a]$
Free coefficients c_i	Computed coefficients $c_i(\theta(a), a)$
Symmetry G (assumed)	Symmetry G (derived from polishing)
G -invariant operators only	All operators (including G -breaking)

The expansions, side by side

Weinberg’s expansion. A Born-layer observable (e.g., a scattering amplitude) at energy E :

$$\mathcal{O}^{\text{Weinberg}}(E) = \underbrace{\mathcal{O}_{\text{LO}}(c_{\text{rel+marg}})}_{\text{relevant + marginal}} + \underbrace{\sum_{n=1}^{\infty} c_n^{(d+n)} \left(\frac{E}{\Lambda}\right)^n}_{\text{irrelevant operators}}. \quad (24.69)$$

The leading order is determined by the relevant and marginal coefficients ($c_{\text{rel+marg}}$: free parameters). The corrections are organised by powers of E/Λ , with one free coefficient $c_n^{(d+n)}$ per independent operator at each order. The expansion is truncated at the desired accuracy.

The lattice-first expansion. The same observable, computed from the engine $H(a)$ with parameters $\theta(a)$:

$$\mathcal{O}^{\text{lattice}}(\theta(a), a, \lambda) = \underbrace{\mathcal{O}_{\text{phys}}(\theta_{\text{rel+marg}}(a))}_{\text{rendering-stable output}} + \underbrace{\sum_{n=1}^{\infty} c_n(\theta(a)) \left(\frac{a}{\lambda}\right)^n}_{\text{lattice artefacts}}. \quad (24.70)$$

The leading order is the rendering-stable Born-layer observable $\mathcal{O}_{\text{phys}}$, determined by the relevant and marginal engine parameters $\theta_{\text{rel+marg}}(a)$ (computed from the Tr log or Feshbach formula). The corrections are lattice artefacts, organised by powers of a/λ , with coefficients $c_n(\theta(a))$ that are *computed functions* of the engine parameters. The rendering-stability condition is: $\mathcal{O} = \mathcal{O}_{\text{phys}} + o(a/\lambda)$.

Where they agree

At leading order ($n = 0$): both expansions give the same observable, determined by the relevant and marginal operators. The structure of the leading-order Lagrangian is the same (it contains all operators with $\Delta \leq d$, consistent with the symmetries). The predictions at leading order agree:

$$\mathcal{O}_{\text{LO}}^{\text{Weinberg}} = \mathcal{O}_{\text{phys}}^{\text{lattice}} + O(a/\lambda). \quad (24.71)$$

The difference is that Weinberg’s LO coefficients are free (measured), while the lattice-first LO coefficients are computed (from $\theta(a)$). If the Weinberg coefficients are determined by matching to the same experiments used to fix $\theta(a)$: the two agree exactly at LO.

At subleading orders ($n \geq 1$): both expansions produce corrections suppressed by the same power of the small parameter ($E/\Lambda \sim a/\lambda$). The operator basis is the same (all G -invariant

operators of dimension $d + n$, if one works within the polished theory). The power counting is the same.

The agreement is not a coincidence. Weinberg’s programme IS the rendering-stability condition, expressed in the language of operators and symmetries. The rendering-stability condition $\mathcal{O}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$ says: the Born-layer output is insensitive to the engine’s resolution. Weinberg’s programme says: the low-energy physics is insensitive to the UV completion. These are the same statement, because $a = 1/\Lambda$ (the engine’s resolution IS the UV cutoff) and $\lambda = 1/E$ (the observation scale IS the inverse energy).

Where they diverge

The divergences are in three places, all traceable to the difference between “free coefficients” and “computed coefficients.”

1. Relations among the coefficients. Weinberg’s expansion at order $(E/\Lambda)^n$ has N_n free coefficients (the number of independent G -invariant operators of dimension $d + n$). These are treated as independent.

The lattice-first expansion has the same N_n coefficients, but they are computed from a smaller number of engine parameters. The engine has N_θ parameters (3 for ϕ^4 , 19 for the SM); the EFT at order n has N_n coefficients. For n large enough: $N_n > N_\theta$, and the lattice-first framework predicts $N_n - N_\theta$ relations among the EFT coefficients.

Example: the gravitational EFT at order R^2 has two independent coefficients c_1 (R^2) and c_2 ($R_{\mu\nu}R^{\mu\nu}$). Weinberg treats them as independent. The lattice-first computation (from the a_2 Seeley–DeWitt coefficient) gives:

$$\frac{c_1}{c_2} = f(N_s, N_f, N_v), \quad (24.72)$$

a definite function of the number of scalars N_s , fermions N_f , and vectors N_v in the engine. This is a prediction: a measurable relation among Weinberg’s “free” coefficients, derivable from the engine.

In ML terms: Weinberg fits N_n parameters independently. The lattice-first framework imposes $N_n - N_\theta$ constraints (like weight sharing in a CNN). The constrained model is more predictive.

2. The G -breaking corrections. Weinberg’s expansion contains only G -invariant operators. The lattice-first expansion contains all operators, including G -breaking lattice artefacts.

At order $(a/\lambda)^n$: the lattice-first expansion has both G -invariant terms (which match Weinberg’s) and G -breaking terms (which Weinberg cannot see). For example, the lattice breaks $O(4)$ to the hypercubic group $H(4)$: operators like $\sum_\mu (\partial_\mu \phi)^4$ (which is $H(4)$ -invariant but not $O(4)$ -invariant) appear in the lattice expansion but not in Weinberg’s.

In practice: the G -breaking terms are suppressed by $(a/\lambda)^2$ relative to the leading G -invariant terms (the lattice $O(4)$ -breaking is an $O(a^2)$ artefact). They are invisible in the rendering regime. But they are there, and the lattice-first framework computes them; Weinberg’s does not.

3. The strong-field regime. Weinberg’s expansion is a power series in E/Λ . It converges (or is at least asymptotically useful) for $E \ll \Lambda$. At $E \sim \Lambda$: the expansion breaks down, every term contributes at $O(1)$, and the EFT is not predictive.

The lattice-first theory does not have this limitation. The engine $H(a)$ is well-defined at all energies $E \leq \pi/a = \Lambda$. The lattice computation (exact diagonalisation, Monte Carlo) gives the Born-layer observable at any E , including $E \sim \Lambda$. The power series $\sum c_n (a/\lambda)^n$ may not converge, but the exact answer (the Tr log, the S -matrix on $\mathcal{F}_a$) is always well-defined.

	Weinberg (E/Λ)	Lattice-first (a/λ)
Expansion parameter	E/Λ (energy / cutoff)	a/λ (lattice / pixel)
Coefficients	Free (N_n per order)	Computed (N_θ engine params determine all)
Relations among c_i	None assumed	$N_n - N_\theta$ predicted at each order
G -breaking terms	Absent (by construction)	Present (lattice artefacts, computable)
Strong field ($E \sim \Lambda$)	Expansion breaks down	Engine computes exactly
Non-perturbative	Not available	Lattice Monte Carlo, exact diag.

The rendering-stability condition IS Weinberg's programme, plus constraints

The rendering-stability condition:

$$\mathcal{O}_{\text{Born}}[\theta(a), a] = \mathcal{O}_{\text{phys}} + o(a/\lambda) \quad (24.73)$$

can be unpacked as follows. Expand $\mathcal{O}$ in the operator basis:

$$\mathcal{O}[\theta(a), a] = \sum_{\Delta_i \leq d} f_i(\theta(a)) \cdot \langle \mathcal{O}_i \rangle_\lambda + \sum_{\Delta_j > d} f_j(\theta(a)) \cdot \left(\frac{a}{\lambda}\right)^{\Delta_j - d} \langle \mathcal{O}_j \rangle_\lambda. \quad (24.74)$$

The first sum (relevant + marginal) gives the leading-order observable, with coefficients $f_i(\theta(a))$ that are functions of the engine parameters. The second sum (irrelevant) gives the corrections, suppressed by $(a/\lambda)^{\Delta_j - d}$.

This IS Weinberg's expansion (24.69), with two additions:

- (i) The coefficients $f_i(\theta(a))$ are *computed*, not free. They satisfy the $N_n - N_\theta$ relations that Weinberg's free parameterisation misses.
- (ii) The G -breaking operators are included in the sum (they are lattice artefacts with their own suppression factors).

If we strip away the constraints (treat the f_i as free, discard the G -breaking terms): we recover Weinberg's programme exactly. The lattice-first framework is Weinberg's programme with the constraints restored.

In one equation.

Weinberg's EFT = rendering stability – coefficient relations – G -breaking corrections.

(24.75)

Weinberg's programme is the rendering-stability condition with the engine's fingerprints erased.

24.10 UV completion, high-energy scattering, and the lattice ceiling

The continuum tradition raises three concerns about quantum field theory at short distances: Does the theory have a UV completion (a well-defined $a \rightarrow 0$ limit)? Is it renormalisable (do finitely many parameters suffice)? Do scattering amplitudes violate unitarity at high energy?

On the lattice, all three dissolve into a single observation: the engine has a finite resolution.

The lattice ceiling

On a lattice with spacing a , momenta are bounded by the Brillouin zone: $|k_\mu| \leq \pi/a$. The maximum centre-of-mass energy squared for a $2 \rightarrow 2$ scattering:

$$s_{\max} \sim \frac{\pi^2 d}{a^2}. \quad (24.76)$$

Scattering at $s > s_{\max}$ is physically impossible: the lattice does not have modes with that much energy. The amplitude $\mathcal{A}(s, t)$ is defined only for $s \leq s_{\max}$.

The S -matrix on $\mathcal{F}_a$ is exactly unitary: $S^\dagger S = \mathbf{1}$. Every Born-layer cross section automatically satisfies the optical theorem, the Froissart bound, partial-wave unitarity, and every other consequence of unitarity. These are not constraints to be checked; they are built into the engine.

Three concerns as continuum extrapolations

Each traditional concern involves pushing the lattice theory beyond its domain:

1. UV completion. The RG flow maps engine parameters $\theta(a) \rightarrow \theta(a')$ as the lattice is refined ($a' < a$). UV completion asks: does the flow have a well-defined limit as $a \rightarrow 0$?

There are three possibilities:

The flow reaches a UV fixed point. $\theta(a) \rightarrow \theta^*$ as $a \rightarrow 0$. The family $\{H(a)\}$ converges to a continuum theory. Example: QCD. The gauge coupling runs to zero (asymptotic freedom):

$$g_s^2(a) \approx \frac{16\pi^2}{b_0 \ln(1/(a\Lambda_{\text{QCD}}))} \xrightarrow{a \rightarrow 0} 0, \quad b_0 = 11 - \frac{2N_f}{3}. \quad (24.77)$$

The Gaussian fixed point $g_s^* = 0$ is UV. QCD has a UV completion.

The flow does not reach a fixed point. $\theta(a)$ diverges or the theory becomes free as $a \rightarrow 0$. Example: ϕ^4 in $d = 4$ (triviality). The coupling runs to zero logarithmically; the continuum limit is a free theory. A nonzero λ_{phys} requires a finite lattice spacing.

The flow reaches a non-Gaussian fixed point. $\theta(a) \rightarrow \theta^* \neq 0$. Example (conjectured): gravity with asymptotic safety (Weinberg, 1979). Evidence from functional RG is suggestive but not conclusive.

In all three cases: the engine at finite a is well-defined and produces rendering-stable Born-layer outputs. UV completion is a question about the rendering's infinite-resolution limit (type B in the classification of Section ??), not about the physics.

2. Renormalizability. Section 24.2 showed: the number of engine parameters needed for rendering stability equals the number of relevant and marginal operators ($\Delta_i \leq d$). In $d = 4$, this count is finite for any theory with finitely many field species. At finite a : the count is always finite (finitely many lattice modes $\Rightarrow$ finitely many independent operators). The distinction between renormalisable and non-renormalisable sharpens only as $a \rightarrow 0$.

3. High-energy unitarity. The traditional argument: compute $\mathcal{A}_{\text{tree}}(s) \sim G_N s$ for graviton exchange. At $s \sim M_P^2$: the amplitude exceeds 1, “violating” partial-wave unitarity. Conclusion: gravity “breaks down” at M_P .

On the lattice: the full amplitude $\mathcal{A}_{\text{full}}(s)$ is unitary at every $s \leq s_{\max}$, because $S^\dagger S = \mathbf{1}$ exactly. The “violation” is a property of the tree-level Taylor approximation, not of the function. This is the same artefact as computing $\sin(x) \approx x$ and concluding $|\sin| > 1$ at $|x| > 1$: the function is bounded; only the approximation violates the bound.

Summary.

Concern	Continuum formulation	Lattice-first status
UV completion	Does the RG flow have a limit as $a \rightarrow 0$?	Type B: question about the rendering. The engine runs at finite a .
Renormalizability	Do finitely many parameters suffice as $a \rightarrow 0$?	At finite a : always yes. The distinction sharpens only as $a \rightarrow 0$.
High-energy unitarity	Does $\mathcal{A}(s)$ violate unitarity at $s \rightarrow \infty$?	Type A: $s \leq s_{\max}$; S is exactly unitary. The “violation” is a Taylor-series artefact.

The operator tower connects all three

The three concerns are manifestations of the same operator classification. In the continuum, the scattering amplitude receives contributions from every operator:

$$\mathcal{A}(s) \sim \sum_i g_i s^{(\Delta_i - d)/2}. \quad (24.78)$$

Each operator of dimension Δ_i contributes a power of s :

Class	s -behavior	Rendering impact	Example
Relevant	$s^{(\Delta - d)/2} \rightarrow 0$: dies	$(a/\lambda)^{d - \Delta}$: grows	Mass term ($\Delta = 2$)
Marginal	$s^0 \cdot \ln^n s$	$\ln(\lambda/a)$	ϕ^4 , F^2 , R^2
Irrelevant	$s^{(\Delta - d)/2} \rightarrow \infty$: grows	$(a/\lambda)^{\Delta - d} \rightarrow 0$: dies	ϕ^6 , F^4 , R^3

The pattern: *high-energy behavior and rendering impact are complementary*. An irrelevant operator is invisible at the rendering scale ($a/\lambda \rightarrow 0$) but dominant at the lattice ceiling ($s \rightarrow s_{\max}$). A relevant operator dominates the rendering but dies at high energy. They are two faces of the same power $(a/\lambda)^{\Delta_i - d} = (s/s_{\max})^{(d - \Delta_i)/2}$.

On the lattice at $s \ll s_{\max}$ (the rendering regime): the irrelevant operators are suppressed by $(s/s_{\max})^{(\Delta_i - d)/2} \sim (a/\lambda)^{\Delta_i - d}$ —the same power that makes them $o(a/\lambda)$ in the rendering-stability condition. They are invisible to the observer.

Near $s \sim s_{\max}$ (the lattice ceiling): the irrelevant operators contribute at $O(1)$, the amplitude reflects the full lattice structure, and the rendering breaks down—the observer is probing the engine at its native resolution. This is not a failure of the theory; it is the edge of the screen.

The unification. Rendering stability (the Born output is a -independent) and high-energy unitarity (the amplitude is bounded) are two views of the same operator classification:

$$\underbrace{(a/\lambda)^{\Delta_i - d} \rightarrow 0}_{\text{rendering stability: irrelevant ops. die}} = \underbrace{(s/s_{\max})^{(\Delta_i - d)/2} \rightarrow 0}_{\text{low-energy unitarity: irrelevant ops. suppressed}} \quad (24.79)$$

The irrelevant operators that wash out of the rendering are the same operators that would cause unitarity violation if extrapolated to $s \rightarrow \infty$. On the lattice, neither extrapolation is taken: a is finite, $s_{\max}$ is finite, and the irrelevant operators are present but harmless at $s \ll s_{\max}$.

The lattice-first punchline

The three concerns—UV completion, renormalizability, high-energy unitarity—are three symptoms of one disease: extrapolating the lattice theory beyond its domain. The lattice cures all three by not extrapolating:

The engine operates at finite a .

The S -matrix is exactly unitary on $\mathcal{F}_a$.

Scattering amplitudes are defined for $s \leq s_{\max}$.

Rendering stability holds with finitely many parameters.

The Born-layer observables (masses, cross sections, decay widths) are finite, unitary, and a -independent to $o(a/\lambda)$.

The physical requirement is rendering stability, not UV completion. The physical domain is $s \leq s_{\max}$, not $s \rightarrow \infty$. The physical S -matrix is unitary exactly, not perturbatively. The “problems” of UV completion, non-renormalisability, and unitarity violation are properties of the continuum extrapolation — the rendering pushed beyond its resolution — not of the engine.

24.11 Mass parameters on the lattice: numerical stability and the hierarchy

The mass parameter $m^2(a)$ is a relevant coupling: its RG eigenvalue $\theta_r = 2 > 0$ means it scales as $1/a^2$. Holding the Born-layer mass m_{phys} fixed while refining the lattice requires $m^2(a) \sim 1/a^2$. This is the hierarchy “problem.”

We have already seen (Section 24.2) that relevant couplings grow under refinement by definition, and (Section 24.10) that the lattice engine operates at finite a where everything is finite. This subsection examines the one genuine concern that survives the lattice-first filter: **numerical stability**.

The perturbative “cancellation” is an artifact

The continuum tradition decomposes the physical mass as:

$$m_{\text{phys}}^2 = \underbrace{m^2(a)}_{\sim 1/a^2} + \underbrace{\Sigma(a)}_{\sim -1/a^2}, \quad (24.80)$$

and notes that two terms of order $1/a^2$ nearly cancel to produce $m_{\text{phys}}^2 \ll 1/a^2$. If $a \sim \ell_P$: a “cancellation to 34 digits.”

This is an artifact of the perturbative decomposition. The function $f(\theta(a), a) = m_{\text{phys}}^2$ is a single smooth function computed non-perturbatively on the lattice (from the exponential decay of $G_E(\tau)$, or by diagonalising $H(a)$). No decomposition into “tree” and “loop” occurs. No large terms appear. The situation is identical to computing $e^{-100} \approx 3.7 \times 10^{-44}$ by its Taylor series: the individual terms reach 10^{40} , but the function itself is one small number. The “cancellation” is in the series, not in the function.

The lattice computes the function directly. It never sees the series.

The real concern: numerical stability near a critical surface

Strip away the perturbative artifact, and the genuine physical content is a statement about the condition number $\kappa = \partial \ln \mathcal{O} / \partial \ln \theta$ (Section ??).

The lattice ϕ^4 theory has a critical surface in its parameter space: the boundary between the symmetric phase ($\langle \phi \rangle = 0$) and the broken phase ($\langle \phi \rangle \neq 0$). The physical Higgs mass is:

$$m_{H,\text{phys}} \sim \frac{1}{a} |m^2(a) - m_c^2|^{1/2}, \quad (24.81)$$

where m_c^2 is the critical value. The condition number:

$$\kappa = \frac{m^2(a)}{m^2(a) - m_c^2} \xrightarrow{m^2(a) \rightarrow m_c^2} \infty. \quad (24.82)$$

A light Higgs ($m_H \cdot a \ll 1$) means the engine operates near the critical surface ($|m^2(a) - m_c^2| \ll 1/a^2$), and $\kappa \gg 1$: a small fractional change in the engine parameter $m^2(a)$ produces a large fractional change in the Born-layer mass m_H .

This is a non-perturbative property of the engine's phase diagram. It does not depend on the Taylor expansion of f . It does not involve any “cancellation.” It is the statement that the function $f(\theta, a)$ has a large derivative near the critical surface.

The question “why is the Higgs light?” becomes, precisely: why does the engine operate close to the critical surface of the Higgs sector?

Symmetry protection as κ control

Not all masses have large κ . A symmetry can keep κ finite—or even zero—by controlling the location of the critical surface relative to the natural parameter values.

Fermion masses: chiral symmetry ($\kappa = O(1)$). For a massless fermion, the Lagrangian has a chiral symmetry $\psi_L \rightarrow e^{i\alpha}\psi_L$, $\psi_R \rightarrow e^{i\beta}\psi_R$. A mass term $m\bar{\psi}\psi$ breaks this symmetry. Quantum corrections to m are proportional to m itself:

$$\delta m_f = \frac{3\alpha}{4\pi} m_f \ln \frac{1}{m_f a} + O(\alpha^2 m_f). \quad (24.83)$$

If $m_f = 0$: it remains zero to all orders. The critical surface ($m_f = 0$) is pinned at the origin by the symmetry. No tuning is needed to stay on it. The condition number $\kappa = O(1)$: the mass responds proportionally to changes in the engine parameter.

On the lattice: this is confirmed directly. Wilson fermions explicitly break chiral symmetry (through the Wilson term $r\bar{\psi}\partial^2\psi$), and they require $O(1/a)$ tuning of the mass parameter to reach the chiral limit—exactly the large- κ behavior expected when the protecting symmetry is broken. Domain-wall and overlap fermions preserve chiral symmetry (up to exponentially small corrections), and they do NOT require power-law tuning. The lattice confirms: mass protection is a property of the symmetry, visible non-perturbatively.

Gauge boson masses: gauge invariance ($\kappa = 0$). A mass term $M^2 A_\mu A^\mu$ breaks gauge invariance. On the lattice with the Wilson plaquette action: gauge invariance is exact, and the Ward identity $p_\mu \Pi^{\mu\nu} = 0$ holds exactly (Section ??). The critical surface for $M \neq 0$ does not exist within the gauge-invariant theory. The condition number is $\kappa = 0$: the photon mass is exactly zero at every a , regardless of the other engine parameters.

Scalar masses: no protection ($\kappa \gg 1$). A scalar mass term $m^2 \phi^2$ does not break any symmetry of the ϕ^4 Lagrangian. The critical surface is at $m_c^2 \neq 0$, and its location is a dynamical quantity (determined by the full phase diagram, not by a symmetry). Near the critical surface: $\kappa \gg 1$. The scalar mass is numerically unstable.

Summary.

Particle	Spin	Protecting symmetry	κ	Critical surface
Fermion	1/2	Chiral	$O(1)$	Pinned at $m = 0$ by symmetry
Gauge boson	1	Gauge invariance	0	Does not exist
Scalar (Higgs)	0	None (in the SM)	$\gg 1$	At $m_c^2 \neq 0$; location is dynamical
Graviton	2	Diff. invariance	0	Does not exist

Does SUSY solve the right problem?

Supersymmetry pairs each boson with a fermion, extending chiral protection to the scalar sector. In SUSY: $\delta m_H^2 \propto (m_{\text{boson}}^2 - m_{\text{fermion}}^2) \ln(1/a)$, proportional to the SUSY-breaking scale rather than $1/a^2$.

From the lattice-first view, this invites scrutiny.

The “problem” that SUSY solves is the large Taylor coefficients in the perturbative expansion of $f(\theta, a)$ —the $\delta m^2 \sim 1/a^2$ one-loop correction. But Section 24.10 established: this correction is a property of the Taylor series, not of the function. On the lattice, f is computed directly. No large terms appear. SUSY cancels terms in an expansion that the lattice never performs.

Meanwhile, SUSY introduces: 32 new particles (squarks, sleptons, gauginos, Higgsinos), 105 new parameters (soft SUSY-breaking terms), and a new algebraic structure (the super-Poincaré algebra)—none of which has been observed. The SUSY-breaking mechanism is unknown and must be added by hand.

The genuine concern is not the Taylor coefficient but the condition number $\kappa \gg 1$. Does SUSY reduce κ ? In a certain sense, yes: SUSY relates the scalar mass to the fermion mass, which is chirally protected ($\kappa = O(1)$). If SUSY is unbroken: $m_H = m_{\tilde{H}}$ exactly, and κ for the scalar inherits the fermion’s stability. But SUSY must be broken (the superpartners are heavier than the Higgs), and the breaking reintroduces $\kappa \gg 1$ at the SUSY-breaking scale. The hierarchy is softened but not eliminated: it is transferred from m_H/M_P to m_H/M_{SUSY} .

The LHC has searched for superpartners at energies up to ~ 2 TeV. None have been found.

Composite Higgs: addressing κ directly

The composite Higgs idea takes a different approach: instead of protecting a fundamental scalar by a new symmetry, it replaces the fundamental scalar with a **fermion bound state**.

In this scenario, the Higgs is not a point particle in the engine; it is a composite object, like the pion in QCD. The pion is a bound state of a quark and an antiquark ($\pi \sim \bar{q}q$), and its mass is protected by chiral symmetry: $m_\pi^2 \propto m_q \cdot \Lambda_{\text{QCD}}$ (the Gell-Mann–Oakes–Renner relation). The pion is light not because of fine-tuning but because it is a pseudo-Nambu–Goldstone boson (pNGB) of the spontaneously broken chiral symmetry.

The composite Higgs replicates this mechanism: the Higgs is a pNGB of a new strong interaction at scale $f \sim \text{TeV}$. New fermions (“technifermions”) are confined by a new gauge force (“technicolor”), and the Higgs emerges as a bound state $H \sim \bar{\psi}_T \psi_T$, with mass protected by the chiral symmetry of the technifermion sector.

From the lattice-first view, the composite Higgs is attractive because it addresses the condition number directly:

The fundamental engine has no scalar. The engine contains only fermions and gauge fields—both of which have protected masses ($\kappa = O(1)$ or $\kappa = 0$). No relevant scalar operator $m^2\phi^2$ appears in the engine. No critical surface for the scalar mass exists.

The Higgs is a rendering-layer object. At the Born-layer observation scale $\lambda \gg 1/f$, the bound state $\bar{\psi}_T\psi_T$ looks like a point scalar, and the effective Lagrangian contains $m_H^2 H^2$. But this “scalar mass” is a Born-layer observable, not an engine parameter. Its value is determined by the technifermion dynamics (which are chirally protected), not by a fine-tuned $m^2(a)$.

The condition number is $O(1)$. The Higgs mass is $m_H \sim gf$, where g is a coupling and f is the symmetry-breaking scale. Both are $O(1)$ in lattice units at the compositeness scale. No large κ . No proximity to a critical surface. The lightness of the Higgs relative to M_P is explained by the ratio $f/M_P \sim 10^{-16}$ —which is a ratio of two dynamical scales (like Λ_{QCD}/M_P), not a fine-tuning of a relevant coupling.

The composite Higgs faces experimental constraints: precision electroweak measurements, flavour-changing neutral currents, and the measured Higgs couplings (which are SM-like to $\sim 10\%$). Modern constructions (“partial compositeness,” “twin Higgs”) address these, but the experimental verdict is not yet in.

The lattice-first summary

Approach	What it does	Does it reduce κ ?	Lattice-first assessment
Perturbative tuning	Adjusts $m^2(a)$ order by order	No ($\kappa \gg 1$ persists)	The “problem”: κ is large near the critical surface
SUSY	Cancels $1/a^2$ in the Taylor series	Partially (transfers hierarchy to m_H/M_{SUSY})	Cures a Taylor-series artifact; introduces 32 particles, 105 parameters
Composite Higgs	Replaces fundamental scalar with pNGB bound state	Yes ($\kappa = O(1)$: chiral protection)	Addresses κ directly; no fundamental scalar in the engine
Anthropic / multi-verse	Accepts $\kappa \gg 1$; invokes selection	No ($\kappa \gg 1$ is accepted)	Consistent with lattice-first view (the engine’s location in parameter space is contingent)

The hierarchy is a fact about the spectrum: $m_H \approx 125 \text{ GeV} \ll M_P \approx 10^{18} \text{ GeV}$. The “problem” is the large condition number $\kappa \gg 1$ of the scalar mass near the critical surface. The perturbative “cancellation to 34 digits” is an artifact. The genuine question is: does the engine architecture explain why κ is moderate, or must $\kappa \gg 1$ be accepted as a contingent fact about the engine’s location in parameter space?

The composite Higgs answers: κ is moderate because the Higgs is not a fundamental scalar (which would have $\kappa \gg 1$) but a fermion bound state (which inherits chiral protection, $\kappa = O(1)$). Whether this is nature’s answer is an experimental question.

25 Lie Groups, Lie Algebras, and Their Equivariant Representations

25.1 Why equivariant representations?

The Lorentz group is a *Lie group*: a group of continuous, differentiable transformations that can be parametrised smoothly by a set of real numbers. A Lorentz boost in the x -direction, for example, is parametrised by the rapidity $\eta \in \mathbb{R}$; a spatial rotation around the z -axis is parametrised by the angle $\theta \in [0, 2\pi)$. Because these parameters vary continuously, we can consider transformations that are infinitesimally close to the identity: a boost by $\eta + d\eta$ differs from a boost by η by an infinitesimal amount. The infinitesimal generators of these transformations — the derivatives of the group elements with respect to the parameters, evaluated at the identity — form the *Lie algebra* of the group. The Lie algebra is a linear vector space with a bilinear bracket operation (the commutator), and it encodes the local structure of the Lie group near the identity. Knowing the Lie algebra is, for most practical purposes, equivalent to knowing the Lie group: every group element can be recovered from the algebra by the matrix exponential $\Lambda = e^{\omega^{\mu\nu} J_{\mu\nu}}$, where $J_{\mu\nu}$ are the generators and $\omega^{\mu\nu}$ are the parameters.

The reason we need to understand Lie groups and their representations is that the field variable $\Phi(x)$ at each spacetime point x must transform *equivariantly* under Lorentz transformations. When we change the spacetime point from x to $x' = \Lambda x$ by a Lorentz transformation Λ , the field variable simultaneously changes from $\Phi(x)$ to $\Phi(x') = S(\Lambda)\Phi(x)$, where $S(\Lambda)$ is a matrix that depends on Λ and on the type of field. The assignment $\Lambda \mapsto S(\Lambda)$ must be a group homomorphism — it must satisfy $S(\Lambda_2\Lambda_1) = S(\Lambda_2)S(\Lambda_1)$ — so that applying two successive Lorentz transformations gives the same result whether we apply them one at a time or together. This is the definition of a *representation* of the Lorentz group: a linear action of the group on the vector space of field values that is consistent with the group structure. For a scalar field, $S(\Lambda) = 1$ (the field value does not rotate); for a vector field, $S(\Lambda) = \Lambda$ (the field value rotates as a four-vector); for a spinor field, $S(\Lambda)$ is a 2×2 or 4×4 matrix in the spinor representation. The complete classification of all possible $S(\Lambda)$ — all possible ways a field can transform equivariantly under the Lorentz group — is the representation theory of the Lorentz group, which we develop in the next section.

Once the equivariant representations are classified, the path to a Lorentz-invariant Lagrangian is clear: assign each field in the theory to a representation $S(\Lambda)$, and contract all Lorentz indices using the invariant tensors of those representations (the Minkowski metric $\eta_{\mu\nu}$, the spinor metric $\epsilon^{ss'}$, the gamma matrices γ_{AB}^μ) so that $\mathcal{L}$ transforms trivially — that is, $S(\Lambda) = 1$ for $\mathcal{L}$ itself. The Lorentz-invariant Lagrangian then provides the starting point for the round trip: Lagrangian $\rightarrow$ Hamiltonian $\rightarrow$ Fock space $\rightarrow$ time slicing $\rightarrow$ path integral $\rightarrow$ Lagrangian. The representation theory developed in the present section is therefore not an abstract mathematical exercise; it is the essential prerequisite for writing down any relativistic quantum field theory. The present section studies equivariant representations of Lie groups and Lie algebras in general, before specialising to the Lorentz group in the next section.

25.2 From e^x to matrix exponentials

The key idea connecting Lie groups to Lie algebras is the matrix exponential. We build up to it in three steps.

Step 1: the real exponential as compound interest. Recall the standard limit:

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n. \quad (25.1)$$

This is the compound-interest formula: starting at 1, apply n small increments of x/n , and in

the limit of infinitely many infinitesimal steps, you get e^x . The Taylor series:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad (25.2)$$

converges for all $x \in \mathbb{R}$.

Step 2: extending to matrices. Replace x by a square matrix X . The Taylor series (25.2) generalises immediately (since matrix powers are well-defined):

$$e^X = \exp(X) = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots, \quad (25.3)$$

and the compound-interest limit also generalises:

$$e^X = \lim_{n \rightarrow \infty} \left(I + \frac{X}{n} \right)^n. \quad (25.4)$$

The interpretation is identical: start at the identity matrix I , apply n small increments of X/n , and in the limit you get e^X . Each step $(I + X/n)$ is a matrix that differs from the identity by a small amount X/n .

Step 3: key properties. The matrix exponential satisfies:

$$e^0 = I, \quad (25.5)$$

$$e^{X+Y} = e^X e^Y \quad \text{if } [X, Y] = XY - YX = 0, \quad (25.6)$$

$$(e^X)^{-1} = e^{-X}, \quad (25.7)$$

$$\det(e^X) = e^{\text{Tr } X}. \quad (25.8)$$

Property (25.6) requires commutativity; for non-commuting matrices the Baker-Campbell-Hausdorff formula applies instead. Property (25.8) is crucial: if $\text{Tr } X = 0$, then $\det(e^X) = 1$.

25.3 Lie groups: continuous symmetry groups

A *Lie group* is a group that is also a smooth manifold: the group operations (multiplication and inversion) are smooth functions of the group parameters. Every element can be reached from the identity by a smooth path, and near the identity, elements look like $I + \epsilon X$ for small ϵ and some matrix X .

Examples.

$GL(n, \mathbb{C})$: all invertible $n \times n$ complex matrices. Dimension $2n^2$ (real).

$U(n)$: unitary matrices, $U^\dagger U = I$. Dimension n^2 .

$SU(n)$: unitary matrices with $\det U = 1$. Dimension $n^2 - 1$.

$SO(n)$: real orthogonal matrices with $\det = +1$. Dimension $n(n-1)/2$.

In physics, the groups that appear most are $SU(2)$, $SU(3)$, $U(1)$, $SL(2, \mathbb{C})$, and $SO(3, 1)$.

One-parameter subgroups. For any matrix X in the Lie algebra (defined below), the map $t \mapsto e^{tX}$ traces a smooth path through the group, passing through the identity at $t = 0$. This path is a *one-parameter subgroup*: a group homomorphism from $(\mathbb{R}, +)$ to the Lie group. Every one-parameter subgroup has this form.

25.4 Lie algebras: infinitesimal generators

The *Lie algebra* $\mathfrak{g}$ of a Lie group G is the tangent space at the identity—the space of all “infinitesimal” group elements. For a matrix group:

$$\mathfrak{g} = \{X : e^{tX} \in G \text{ for all } t \in \mathbb{R}\}. \quad (25.9)$$

The generators T^a are a basis for $\mathfrak{g}$, and any element near the identity can be written $g = e^{i\theta^a T^a}$ for small θ^a . (The factor of i is conventional in physics to make T^a Hermitian when G is unitary.)

The Lie bracket and structure constants. The Lie algebra is not just a vector space—it has a bilinear operation, the *Lie bracket* (commutator):

$$[T^a, T^b] = if^{abc}T^c, \quad (25.10)$$

where f^{abc} are the *structure constants* of the algebra. The commutator of two generators is again a generator (the Lie algebra is closed under commutation). The structure constants satisfy $f^{abc} = -f^{bac}$ (antisymmetry) and the Jacobi identity $f^{ade}f^{bcd} + f^{bde}f^{cad} + f^{cde}f^{abd} = 0$.

Why the commutator? Consider two small group transformations $e^{\epsilon X}$ and $e^{\delta Y}$. The combined transformation $e^{\epsilon X}e^{\delta Y}e^{-\epsilon X}e^{-\delta Y}$ measures how much X and Y fail to commute. Expanding to second order:

$$e^{\epsilon X}e^{\delta Y}e^{-\epsilon X}e^{-\delta Y} = I + \epsilon\delta[X, Y] + O(\epsilon^2, \delta^2). \quad (25.11)$$

The commutator $[X, Y]$ is the infinitesimal generator of this “commutator loop”—it measures the curvature of the group manifold. The structure constants encode how the generators interact; they completely characterise the algebraic structure of the group near the identity.

Lie algebra vs. Lie group. This distinction is crucial and often blurred:

	Lie group G	Lie algebra $\mathfrak{g}$
What it is	Manifold + group structure	Vector space + Lie bracket
Elements	Group elements $g \in G$	Generators $X \in \mathfrak{g}$
Near identity	$g \approx I + \epsilon X$	$X = \text{tangent vector at } I$
Global structure	Yes (topology matters)	No (only local info)
Exponential	$g = e^X$ (not always surjective)	$X = \log g$ (near identity)

The Lie algebra captures the *local* structure of the Lie group near the identity. Two different Lie groups can share the same Lie algebra if they differ only in their global topology—we will see this with $SU(2)$ and $SO(3)$ below.

25.5 $SU(2)$ and $SO(3)$: the same algebra, different groups

The group $SU(2)$. $SU(2)$ consists of 2×2 unitary matrices with determinant 1:

$$SU(2) = \left\{ U = \begin{pmatrix} a & -b^* \\ b & a^* \end{pmatrix} : a, b \in \mathbb{C}, |a|^2 + |b|^2 = 1 \right\}. \quad (25.12)$$

As a manifold, $SU(2) \cong S^3$ (the 3-sphere). It is simply connected (every loop can be contracted to a point). Dimension 3.

The generators (basis of $\mathfrak{su}(2)$) are $T^a = \sigma^a/2$ where σ^a are the Pauli matrices:

$$T^1 = \frac{\sigma^1}{2} = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad T^2 = \frac{\sigma^2}{2} = \frac{1}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad T^3 = \frac{\sigma^3}{2} = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (25.13)$$

They satisfy the Lie algebra:

$$[T^a, T^b] = i\epsilon^{abc}T^c, \quad (25.14)$$

with structure constants $f^{abc} = \epsilon^{abc}$ (Levi-Civita symbol).

A general $SU(2)$ element in the direction $\hat{n} = (n_1, n_2, n_3)$ with angle θ is:

$$U = e^{i\theta\hat{n}\cdot\vec{T}} = e^{i\frac{\theta}{2}\hat{n}\cdot\vec{\sigma}} = I \cos \frac{\theta}{2} + i(\hat{n} \cdot \vec{\sigma}) \sin \frac{\theta}{2}. \quad (25.15)$$

The group $SO(3)$. $SO(3)$ consists of 3×3 real orthogonal matrices with determinant 1: the rotation matrices in three dimensions. As a manifold, $SO(3) \cong \mathbb{RP}^3$ (real projective space), which is *not* simply connected. The fundamental group is $\pi_1(SO(3)) = \mathbb{Z}/2$: there are loops that cannot be contracted to a point. Dimension 3.

The generators of $SO(3)$ (angular momentum matrices in 3D) are:

$$J^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad J^2 = \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{pmatrix}, \quad J^3 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (25.16)$$

satisfying exactly the same Lie algebra:

$$[J^a, J^b] = i\epsilon^{abc}J^c. \quad (25.17)$$

Same algebra, different groups. $\mathfrak{su}(2) \cong \mathfrak{so}(3)$: both algebras are defined by $[T^a, T^b] = i\epsilon^{abc}T^c$ with structure constants ϵ^{abc} . The generators look different (2×2 vs 3×3 matrices) but they obey the same commutation relations.

Yet the *groups* differ:

	$SU(2)$	$SO(3)$
Manifold	S^3 (3-sphere)	$\mathbb{RP}^3$
Simply connected?	Yes	No
Fundamental group	$\{e\}$	$\mathbb{Z}/2$
360° rotation	$e^{i\pi\hat{n}\cdot\vec{T}} = -I$	$e^{i\pi\hat{n}\cdot\vec{J}} = +I$
720° rotation	$+I$	$+I$

The key difference: in $SU(2)$, a 360° rotation gives $-I$ (not the identity), while in $SO(3)$ it gives $+I$. $SU(2)$ is the *universal cover* of $SO(3)$: there is a 2-to-1 group homomorphism $\phi: SU(2) \rightarrow SO(3)$ with kernel $\{+I, -I\}$. Both $+U$ and $-U$ in $SU(2)$ map to the same rotation in $SO(3)$.

Example. The 2-to-1 map $\phi: SU(2) \rightarrow SO(3)$.

For $U \in SU(2)$, define the map $\phi(U): \mathbb{R}^3 \rightarrow \mathbb{R}^3$ by its action on vectors $\mathbf{v} = (v_1, v_2, v_3)$:

$$\phi(U): v^a \sigma^a \mapsto U(v^a \sigma^a)U^\dagger.$$

Since $U^\dagger U = I$ and $\det(v^a \sigma^a) = -|\mathbf{v}|^2$ is preserved, $\phi(U)$ is a rotation. Both U and $-U$ give the same rotation since $(-U)X(-U)^\dagger = UXU^\dagger$. The map is 2-to-1.

25.6 Representations of Lie algebras and Lie groups

A *representation* of a Lie group G is a smooth group homomorphism $\rho : G \rightarrow GL(V)$ (the group of invertible linear maps on a vector space V). It assigns to each group element a linear transformation of V , consistently with the group law: $\rho(gh) = \rho(g)\rho(h)$.

A *representation of the Lie algebra* $\mathfrak{g}$ is a linear map $d\rho : \mathfrak{g} \rightarrow \mathfrak{gl}(V)$ that preserves the Lie bracket: $d\rho([X, Y]) = [d\rho(X), d\rho(Y)]$.

The two are related by exponentiation:

$$\rho(e^X) = e^{d\rho(X)}. \quad (25.18)$$

Given a Lie algebra representation $d\rho$, exponentiating gives a Lie group representation (at least locally; globally one must check for topological subtleties).

25.7 Representations of $\mathfrak{su}(2)$: the spin- j family

The representations of $\mathfrak{su}(2)$ (and hence of $SU(2)$) are completely classified. We derive them by finding all ways to represent the algebra $[J^a, J^b] = i\epsilon^{abc}J^c$ on a vector space.

Raising and lowering operators. Define:

$$J_{\pm} = J^1 \pm iJ^2, \quad J_3 = J^3. \quad (25.19)$$

The algebra (25.14) becomes:

$$[J_3, J_{\pm}] = \pm J_{\pm}, \quad [J_3, J_-] = -J_-, \quad [J_+, J_-] = 2J_3. \quad (25.20)$$

J_3 is diagonal; J_+ raises its eigenvalue by 1; J_- lowers by 1.

The highest-weight construction. In any finite-dimensional irreducible representation (irrep), there must be a vector $|j, j\rangle$ annihilated by J_+ (a highest-weight vector):

$$J_+ |j, j\rangle = 0, \quad J_3 |j, j\rangle = j |j, j\rangle. \quad (25.21)$$

Applying J_- repeatedly:

$$|j, m\rangle \propto (J_-)^{j-m} |j, j\rangle, \quad m = j, j-1, \dots, -j. \quad (25.22)$$

The representation terminates (is finite-dimensional) when $J_- |j, -j\rangle = 0$. This forces $2j$ to be a non-negative integer, so $j \in \{0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots\}$. The dimension of the representation is $2j+1$.

Explicit matrix form. In the basis $\{|j, m\rangle : m = -j, \dots, +j\}$, the generators act as:

$$J_3 |j, m\rangle = m |j, m\rangle, \quad (25.23)$$

$$J_{\pm} |j, m\rangle = \sqrt{j(j+1) - m(m \pm 1)} |j, m \pm 1\rangle. \quad (25.24)$$

The matrices $D^{(j)}(T^a)$ are the $(2j+1) \times (2j+1)$ matrices of the spin- j representation.

Example. Spin- $\frac{1}{2}$ ($j = \frac{1}{2}$): the fundamental representation.

Basis: $\{|\frac{1}{2}, +\frac{1}{2}\rangle, |\frac{1}{2}, -\frac{1}{2}\rangle\}$.

$$\begin{aligned} J_3 &= \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \frac{\sigma^3}{2}, \\ J_+ &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad J_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \\ J^1 &= \frac{J_+ + J_-}{2} = \frac{\sigma^1}{2}, \quad J^2 = \frac{J_+ - J_-}{2i} = \frac{\sigma^2}{2}. \end{aligned}$$

This is exactly the $\mathfrak{su}(2)$ generators (25.13)—the fundamental representation on $\mathbb{C}^2$.

Example. Spin-1 ($j = 1$): the adjoint / 3-vector representation.

Basis: $\{|1, +1\rangle, |1, 0\rangle, |1, -1\rangle\}$.

$$J_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad J_+ = \sqrt{2} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad J_- = \sqrt{2} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The 3×3 generators J^a for $SO(3)$ (25.16) are exactly the spin-1 representation matrices. This is the adjoint representation of $SU(2)$ (and the defining representation of $SO(3)$).

The spin- j representation table.

j	$\dim = 2j + 1$	Name	$SU(2)?$	$SO(3)?$
0	1	Scalar (trivial)	Yes	Yes
$\frac{1}{2}$	2	Spinor (fundamental)	Yes	No
1	3	Vector (adjoint)	Yes	Yes
$\frac{3}{2}$	4	Rarita-Schwinger	Yes	No
2	5	Symmetric tensor	Yes	Yes
$\frac{5}{2}$	6	—	Yes	No
$j \in \mathbb{Z}$	$2j + 1$ odd	Integer spin	Yes	Yes
$j \in \mathbb{Z} + \frac{1}{2}$	$2j + 1$ even	Half-integer spin	Yes	No

25.8 Why half-integer spins do not represent $SO(3)$

The algebra $\mathfrak{su}(2) \cong \mathfrak{so}(3)$ admits all half-integer and integer spin representations. But for the group $SO(3)$, only integer spins are valid representations. Why?

The obstruction: the 360° rotation. In $SO(3)$, a 360° rotation about any axis is the identity transformation (it maps every vector in $\mathbb{R}^3$ to itself). So any representation ρ of $SO(3)$ must satisfy:

$$\rho(R_{360^\circ}) = I, \quad (25.25)$$

where R_{360° is the 360° rotation matrix (the identity in $SO(3)$).

In the spin- j representation, the element $R_{2\pi\hat{n}} \in SU(2)$ corresponding to a 360° rotation is:

$$e^{i \cdot 2\pi\hat{n} \cdot \vec{J}} = e^{i \cdot 2\pi j} \cdot I_{2j+1} = (-1)^{2j} I_{2j+1}. \quad (25.26)$$

For integer j : $(-1)^{2j} = +1$. The constraint (25.25) is satisfied. The spin- j representation descends to a representation of $SO(3)$.

For half-integer j : $(-1)^{2j} = -1$. The spin- j representation sends the identity element of $SO(3)$ to $-I$, not $+I$. This is not a valid group homomorphism. Half-integer spin representations are representations of $SU(2)$ but not of $SO(3)$.

Projective representations. Half-integer spins *do* represent $SO(3)$ projectively: as maps $\rho : SO(3) \rightarrow PGL(V)$ (into the projective general linear group, where matrices differing by a scalar are identified). In quantum mechanics, states are rays in Hilbert space (equivalence classes of vectors differing by a phase), so projective representations are physically acceptable. The electron is a spin- $\frac{1}{2}$ projective representation of $SO(3)$ —equivalently, a genuine representation of the universal cover $SU(2)$.

25.9 Infinite-dimensional representations and function spaces

The representations above are finite-dimensional ($\dim V = 2j + 1$). But the natural action of $SU(2)$ and $SO(3)$ on function spaces is infinite-dimensional—and it decomposes into a direct sum of the finite-dimensional irreps.

$L^2(S^2)$: functions on the 2-sphere. The rotation group $SO(3)$ acts on functions $f : S^2 \rightarrow \mathbb{C}$ by:

$$[\rho(R)f](\hat{n}) = f(R^{-1}\hat{n}), \quad (25.27)$$

rotating the argument. This defines an infinite-dimensional representation on the Hilbert space $L^2(S^2)$ (square-integrable functions on the sphere, with inner product $\langle f, g \rangle = \int_{S^2} f^* g d\Omega$).

Spherical harmonics as basis. The *spherical harmonics* $Y_\ell^m(\theta, \phi)$ (with $\ell \in \{0, 1, 2, \dots\}$ and $m \in \{-\ell, \dots, +\ell\}$) are the eigenfunctions of $J^2 = (J^1)^2 + (J^2)^2 + (J^3)^2$ and J_3 :

$$J^2 Y_\ell^m = \ell(\ell + 1) Y_\ell^m, \quad J_3 Y_\ell^m = m Y_\ell^m. \quad (25.28)$$

They span $L^2(S^2)$: every square-integrable function on the sphere expands as $f = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} c_{\ell m} Y_\ell^m$. The decomposition of $L^2(S^2)$ into irreps of $SO(3)$ is:

$$L^2(S^2) = \bigoplus_{\ell=0}^{\infty} V_\ell, \quad (25.29)$$

where V_ℓ is the spin- ℓ irrep spanned by $\{Y_\ell^m : m = -\ell, \dots, +\ell\}$.

Note that only *integer* ℓ appears in (25.29): these are $SO(3)$ representations (not $SU(2)$ -only ones). The spherical harmonics Y_ℓ^m are explicitly:

$$Y_0^0 = \frac{1}{\sqrt{4\pi}}, \quad Y_1^0 = \sqrt{\frac{3}{4\pi}} \cos \theta, \quad Y_1^{\pm 1} = \mp \sqrt{\frac{3}{8\pi}} \sin \theta e^{\pm i\phi}, \quad \dots \quad (25.30)$$

These are polynomials in $(\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta) = (x, y, z)/r$: the homogeneous polynomials of degree ℓ restricted to the sphere.

$L^2(S^3)$: functions on the 3-sphere and $SU(2)$. For $SU(2) \cong S^3$, the analogous decomposition of $L^2(SU(2))$ (square-integrable functions on $SU(2)$ under the Haar measure) is:

$$L^2(SU(2)) = \bigoplus_{j=0, \frac{1}{2}, 1, \frac{3}{2}, \dots} V_j^L \otimes V_j^R, \quad (25.31)$$

where V_j^L and V_j^R are copies of the spin- j irrep acting on the left and right of $SU(2)$. Now *all* half-integer j appear, because $SU(2)$ is the correct group for which they are valid representations.

The analogue of spherical harmonics for $SU(2)$ are the *Wigner D-matrices* $D_{mm'}^{(j)}(U)$: matrix elements of the spin- j representation as functions on the group. They satisfy orthogonality relations analogous to the spherical harmonic orthogonality $\int Y_\ell^{m*} Y_{\ell'}^{m'} d\Omega = \delta_{\ell\ell'} \delta_{mm'}$.

The hydrogen atom: energy eigenstates as $SO(3)$ irreps. The most familiar physical example is the hydrogen atom. The Hilbert space decomposes as:

$$\mathcal{H}_H = \bigoplus_{n=1}^{\infty} \bigoplus_{\ell=0}^{n-1} V_\ell, \quad (25.32)$$

where V_ℓ is the $SO(3)$ irrep of spin ℓ spanned by $\{Y_\ell^m : m = -\ell, \dots, +\ell\}$. The quantum numbers n, ℓ, m of the hydrogen orbitals are precisely: n the energy level, ℓ the $SO(3)$ irrep label

(total angular momentum), and m the weight of the state within the irrep (angular momentum projection). This is the representation theory of $SO(3)$ appearing directly in chemistry.

Spin- $\frac{1}{2}$ particles need $SU(2)$, not $SO(3)$. The electron has spin $j = \frac{1}{2}$: a two-component spinor ψ^m ($m \in \{+\frac{1}{2}, -\frac{1}{2}\}$) transforming as the fundamental representation of $SU(2)$. This is not a representation of $SO(3)$ (a 360° rotation gives $-\psi$, not $+\psi$). The correct symmetry group for spin- $\frac{1}{2}$ particles is $SU(2)$, and this is why the Dirac equation requires $SL(2, \mathbb{C})$ (the relativistic cover of $SU(2)$) rather than the Lorentz group $SO^+(3, 1)$ directly.

Space	Decomposes into	Physical example
$L^2(S^2)$	$\bigoplus_{\ell=0}^{\infty} V_\ell$ ($SO(3)$ reps, integer only)	Orbital wavefunctions
$L^2(SU(2))$	$\bigoplus_{j=0, \frac{1}{2}, \dots} V_j^L \otimes V_j^R$ (all j)	Wigner D-matrices
$\mathcal{H}_H$	$\bigoplus_{n, \ell} V_\ell$ ($SO(3)$)	Hydrogen atom orbitals
$\mathcal{H}_{e^-}$	spin- $\frac{1}{2}$ irrep of $SU(2)$	Electron spin
Fock space $\mathcal{F}$	direct sum over all particle configurations	Standard Model

25.10 Connection to equivariant representation learning

The language of Lie group representations is not foreign to machine learning practitioners—it is the formal version of something ML already does informally. In representation learning, one maps objects (images, molecules, tokens) to vectors that are acted on by transformations in a structured way. Lie group representations make this structure exact.

The basic setup: nouns and verbs

Consider a set of objects $\mathcal{X}$ (points on a sphere, molecules, states in a physical system) and a group G of transformations acting on them: $x \mapsto g \cdot x$ for $g \in G$, $x \in \mathcal{X}$.

A *representation* assigns to each object x an embedding vector $v(x) \in V$, and to each group element g a matrix $M(g) \in GL(V)$, such that the embedding intertwines the group action:

$$v(g \cdot x) = M(g) v(x). \quad (25.33)$$

In the language of ML: $v(x)$ is the *embedding* or *representation* of the object x (a noun); $M(g)$ is the *representation* of the transformation g (a verb). The equation (25.33) says: *transforming the object and then embedding gives the same result as embedding and then applying the matrix*. The embedding and the transformation commute.

Commutative diagram: objects and embeddings.

$$\begin{array}{ccc}
 x & \xrightarrow{g \cdot} & g \cdot x \\
 \downarrow v & & \downarrow v \\
 v(x) & \xrightarrow{M(g)} & M(g) v(x)
 \end{array}$$

The top arrow is the action in the *native space* $\mathcal{X}$; the bottom arrow is the action in the *representation space* V . The diagram commutes if and only if v is equivariant with respect to G . This is the definition of an equivariant map, and it is exactly what a representation of G provides.

ML reading. In a graph neural network that is equivariant to permutations of nodes, the embedding of a permuted graph equals the permuted embedding of the original graph. In an

SE(3)-equivariant network for molecular geometry, rotating a molecule and then embedding gives the same vector as embedding and then applying the rotation matrix. In both cases, the network implements (25.33) for the appropriate group G (permutations, rotations). The representation matrices $M(g)$ are the group's action on feature space.

Infinitesimal version: Lie algebra and vector fields

For a Lie group, the transformation $g = e^{\epsilon X}$ (small ϵ) acts on $\mathcal{X}$ as:

$$g \cdot x = e^{\epsilon X} \cdot x \approx x + \epsilon \xi_X(x), \quad (25.34)$$

where $\xi_X(x) = \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} (e^{\epsilon X} \cdot x)$ is the *vector field* on $\mathcal{X}$ generated by X . The vector field tells you the infinitesimal direction of motion under the group action.

The corresponding action on the embedding is:

$$v(g \cdot x) = M(e^{\epsilon X}) v(x) \approx (I + \epsilon d\rho(X)) v(x) = v(x) + \epsilon d\rho(X) v(x), \quad (25.35)$$

where $d\rho(X)$ is the Lie algebra representation matrix (the infinitesimal generator in V).

Commutative diagram: infinitesimal version.

$$\begin{array}{ccc} x & \xrightarrow{\xi_X(x) dt} & x + \xi_X(x) dt \\ \downarrow v & & \downarrow v \\ v(x) & \xrightarrow{d\rho(X) dt} & v(x) + d\rho(X) v(x) dt \end{array}$$

The top arrow is the infinitesimal flow on $\mathcal{X}$ generated by X ; the bottom arrow is the corresponding infinitesimal action on V via the Lie algebra generator $d\rho(X)$. Commutativity of this diagram (to first order in dt) is exactly the equivariance condition (25.33) differentiated at the identity.

The Lie algebra acts as a differential operator. The generator X acts on *functions* on $\mathcal{X}$ via the vector field ξ_X :

$$(Xf)(x) = \xi_X(x) \cdot \nabla f(x) = \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} f(e^{\epsilon X} \cdot x). \quad (25.36)$$

This is the *differential operator representation* of the Lie algebra on function space. The generator does not act by a matrix; it acts by a differential operator. The representation matrices $d\rho(X)$ of the finite-dimensional irreps are the restrictions of these differential operators to finite-dimensional invariant subspaces.

Example. $SO(3)$ acting on functions on the sphere.

$SO(3)$ acts on $S^2 \subset \mathbb{R}^3$ by rotations $x \mapsto Rx$. The generator J^3 (rotation about the z -axis) acts on functions by:

$$(J^3 f)(\theta, \phi) = \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} f(\theta, \phi - \epsilon) = -\frac{\partial f}{\partial \phi}.$$

The spherical harmonics are eigenfunctions of $-iJ^3 = -i\partial_\phi$:

$$-i \frac{\partial}{\partial \phi} Y_\ell^m = m Y_\ell^m.$$

The eigenvalue m is the weight of the irrep V_ℓ in which Y_ℓ^m lives. The differential operator J^3 on $L^2(S^2)$ restricts to the finite-dimensional matrix $J_3 = \text{diag}(\ell, \ell-1, \dots, -\ell)$ on each invariant subspace $V_\ell = \text{span}\{Y_\ell^m : m = -\ell, \dots, \ell\}$.

Embeddings, equivariance, and basis functions

The connection to representation learning is now complete. The spherical harmonics Y_ℓ^m are not just abstract eigenfunctions — they are an *equivariant embedding* of the sphere S^2 into function space:

$$v_\ell : S^2 \rightarrow V_\ell \cong \mathbb{C}^{2\ell+1}, \quad \hat{n} \mapsto (Y_\ell^{-\ell}(\hat{n}), \dots, Y_\ell^\ell(\hat{n})). \quad (25.37)$$

For each ℓ , this is an equivariant map: rotating $\hat{n}$ by R and then computing Y_ℓ^m gives the same result as computing $Y_\ell^m(\hat{n})$ and then acting with the spin- ℓ rotation matrix $D^{(\ell)}(R)$:

$$Y_\ell^m(R^{-1}\hat{n}) = \sum_{m'=-\ell}^{\ell} D^{(\ell)}(R)^m_{m'} Y_\ell^{m'}(\hat{n}). \quad (25.38)$$

This is exactly (25.33) for $\mathcal{X} = S^2$, $G = SO(3)$, $v = v_\ell$, $M(g) = D^{(\ell)}(R)$. The spherical harmonics are the equivariant basis functions for functions on the sphere, exactly as token embeddings are the basis functions for language models.

From functions to physics. In quantum mechanics, the wavefunction $\psi(\mathbf{x})$ is an element of $L^2(\mathbb{R}^3)$. The decomposition of $L^2(S^2)$ into spherical harmonics is the angular part of this decomposition. The orbital quantum numbers (ℓ, m) label the irrep V_ℓ and the weight within it — they are the equivariant embedding coordinates. The radial wavefunction $R_{n\ell}(r)$ is the “content” at fixed equivariant coordinates (ℓ, m) .

In QFT, the field $\Phi(x)$ is a map from spacetime to a representation space V of the Lorentz group. The mode expansion in terms of plane waves $e^{ik \cdot x}$ is the QFT analogue of the spherical harmonic expansion: the plane waves are the equivariant basis functions for the translation group, and the creation/annihilation operators $a^\dagger(\mathbf{k})$ are the equivariant embedding coordinates (momentum-space labels of the Fock basis).

The unified picture.

Domain	Objects $\mathcal{X}$	Group G	Equivariant basis
Sphere S^2	Points $\hat{n}$	$SO(3)$	Spherical harmonics Y_ℓ^m
$\mathbb{R}^3$	Points $\mathbf{x}$	Translations	Plane waves $e^{i\mathbf{k} \cdot \mathbf{x}}$
Spacetime	Points x^μ	Lorentz + transl	Plane waves $e^{ik_\mu x^\mu}$
Molecules	Atom positions	$SE(3)$	Spherical harmonics on bonds
Language	Tokens	(no symmetry)	Learned embeddings
Fock space	Particle configs	G_{SM}	Momentum-spin basis $ \mathbf{k}, s, \dots\rangle$

In each row, the equivariant basis functions are the analogues of spherical harmonics for that domain and group. The representation matrices $M(g) = D^{(j)}(g)$ tell you how the basis transforms. The embedding $v(x)$ expresses the object in terms of these basis functions.

Nouns and verbs, revisited. In the commutative diagram, $v(x)$ is the embedding of an object (a *noun*: a point, a molecule, a particle configuration). $M(g)$ is the representation of a transformation (a *verb*: a rotation, a boost, a gauge transformation). The representation $\rho : G \rightarrow GL(V)$ assigns a matrix verb $M(g)$ to every abstract verb $g \in G$. The equivariance condition (25.33) is the requirement that verbs act consistently on noun embeddings: acting on

the noun and then embedding gives the same result as embedding and then applying the matrix verb.

In the Standard Model: the noun is a particle configuration $|n; \mathbf{k}_1, I_1, \dots\rangle$; the verb is a Lorentz transformation Λ or a gauge transformation $U(x)$; the representation matrices $D^{(j)}(L)$, U^α_β , V^i_j are the matrix verbs that act on the spinor and gauge indices of the field. The Schrödinger RNN update $h_t \rightarrow e^{-iHdt} h_t$ is a time-evolution verb applied to the noun h_t . The requirement that H commutes with all symmetry transformations (Noether's theorem) is exactly the requirement that time evolution is an equivariant map with respect to the symmetry group.

25.11 3D vectors as 2×2 matrices: the adjoint action

We have seen that $SU(2)$ is the double cover of $SO(3)$: every rotation in 3D corresponds to two elements of $SU(2)$, differing by a sign. This relationship becomes completely explicit when we represent 3D vectors not as columns in $\mathbb{R}^3$ but as Hermitian traceless 2×2 complex matrices. The Pauli matrices provide the bridge, and the resulting picture illuminates the meaning of the double cover, gives the cleanest example of an equivariant representation, and foreshadows the Lorentz group.

Encoding a 3D vector as a 2×2 matrix

Given a vector $\mathbf{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$, define the associated Hermitian traceless matrix:

$$V = \mathbf{v} \cdot \boldsymbol{\sigma} = v_1 \sigma_1 + v_2 \sigma_2 + v_3 \sigma_3 = \begin{pmatrix} v_3 & v_1 - iv_2 \\ v_1 + iv_2 & -v_3 \end{pmatrix}. \quad (25.39)$$

This map $\mathbf{v} \mapsto V$ is a linear isomorphism between $\mathbb{R}^3$ and the space of Hermitian traceless 2×2 matrices:

$$\mathbb{R}^3 \xrightarrow{\sim} \mathfrak{h} = \{M \in M_2(\mathbb{C}) : M = M^\dagger, \text{Tr}(M) = 0\}, \quad (25.40)$$

with inverse $v_k = \frac{1}{2} \text{Tr}(\sigma_k V)$. The Euclidean norm is encoded in the determinant:

$$\det V = -v_3^2 - (v_1^2 + v_2^2) = -|\mathbf{v}|^2. \quad (25.41)$$

The length of the vector is the square root of minus the determinant of its matrix representative.

Rotations act by conjugation

Let $U \in SU(2)$. Define the *adjoint action* on $\mathfrak{h}$:

$$\text{Ad}_U : V \mapsto UVU^\dagger. \quad (25.42)$$

Three facts confirm this is a rotation:

1. **Stays in $\mathfrak{h}$:** $(UVU^\dagger)^\dagger = UV^\dagger U^\dagger = UVU^\dagger$ and $\text{Tr}(UVU^\dagger) = \text{Tr}(V) = 0$.
2. **Preserves norm:** $\det(UVU^\dagger) = \det U \cdot \det V \cdot \det U^\dagger = 1 \cdot \det V \cdot 1 = \det V$, so $|\mathbf{v}|$ is unchanged.
3. **Group homomorphism:** $\text{Ad}_{UV} = \text{Ad}_U \circ \text{Ad}_V$.

Since Ad_U is a linear norm-preserving map on $\mathbb{R}^3$ and $SU(2)$ is connected, $\text{Ad}_U \in SO(3)$. We have a group homomorphism:

$$\Phi : SU(2) \longrightarrow SO(3), \quad \Phi(U)(\mathbf{v}) = \text{vector corresponding to } UVU^\dagger. \quad (25.43)$$

Explicit example. Take $U(\theta) = e^{i\theta\sigma_3/2} = \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix}$. Then:

$$U(\theta)\sigma_1U(\theta)^\dagger = \cos\theta\sigma_1 - \sin\theta\sigma_2, \quad U(\theta)\sigma_2U(\theta)^\dagger = \sin\theta\sigma_1 + \cos\theta\sigma_2, \quad U(\theta)\sigma_3U(\theta)^\dagger = \sigma_3. \quad (25.44)$$

The adjoint action of $e^{i\theta\sigma_3/2}$ is a rotation by angle θ around the z -axis. Note that the $SU(2)$ parameter is $\theta/2$: a rotation by θ in $SO(3)$ corresponds to an $SU(2)$ element at half the angle.

Commutative diagram: equivariance

The adjoint action is an equivariant map. On the left side of the diagram, $SU(2)$ acts on a vector $\mathbf{v}$ via the rotation $\Phi(U) \in SO(3)$. On the right side, $SU(2)$ acts on the matrix V via conjugation. The encoding $\mathbf{v} \mapsto V$ intertwines these two actions:

$$\begin{array}{ccc} \mathbf{v} \in \mathbb{R}^3 & \xrightarrow{\text{encode}} & V = \mathbf{v} \cdot \boldsymbol{\sigma} \in \mathfrak{h} \\ \Phi(U) \text{ (rotation)} \downarrow & & \downarrow \text{Ad}_U \text{ (conjugation)} \\ \mathbf{v}' = \Phi(U)\mathbf{v} \in \mathbb{R}^3 & \xrightarrow{\text{encode}} & V' = UVU^\dagger \in \mathfrak{h} \end{array}$$

The diagram commutes: rotating first and then encoding gives the same result as encoding first and then conjugating. In the language of representations, the encoding $\mathbf{v} \mapsto V$ is an *intertwiner* between:

- The standard representation of $SO(3)$ on $\mathbb{R}^3$ (acting by $\Phi(U)$ on column vectors), and
- The adjoint representation of $SU(2)$ on $\mathfrak{h}$ (acting by $U(\cdot)U^\dagger$ on matrices).

These two representations are isomorphic as representations of $SU(2)$. This is exactly the equivariance condition we require of an equivariant neural network: the feature map (encoding) commutes with the group action. Here the “input space” is $\mathbb{R}^3$, the “feature space” is $\mathfrak{h}$, and the group is $SU(2)$ acting by two different but equivalent representations.

The double cover: U and $-U$ give the same rotation

The kernel of Φ is the set of $U \in SU(2)$ that commute with every element of $\mathfrak{h}$, i.e. every Pauli matrix. The only 2×2 matrices that commute with all three Pauli matrices are multiples of the identity, and the only multiples in $SU(2)$ are $\pm I$:

$$\ker \Phi = \{I, -I\}. \quad (25.45)$$

Therefore $\Phi(U) = \Phi(-U)$ for all U : the two elements U and $-U$ of $SU(2)$ produce exactly the same rotation in $SO(3)$. The map Φ is two-to-one.

Tracking U through a 360° rotation. Consider the path in $SU(2)$:

$$U(\theta) = e^{i(\theta/2)\hat{n} \cdot \boldsymbol{\sigma}}, \quad \theta \in [0, 4\pi], \quad (25.46)$$

for a fixed unit axis $\hat{n}$. The corresponding rotation in $SO(3)$ is $R(\theta) = \Phi(U(\theta))$, a rotation by angle θ around $\hat{n}$.

$$\begin{array}{ccccccc}
U = I & \xrightarrow{0 \rightarrow \pi} & U = i\hat{n} \cdot \boldsymbol{\sigma} & \xrightarrow{\pi \rightarrow 2\pi} & U = -I & \xrightarrow{2\pi \rightarrow 4\pi} & U = I \\
\downarrow \Phi & & \downarrow \Phi & & \downarrow \Phi & & \downarrow \Phi \\
R = I & \xrightarrow{\text{grows}} & R = R_{\hat{n}}(\pi) & \xrightarrow{\text{returns}} & R = I & \xleftarrow{\text{(stays)}} & R = I
\end{array}$$

$\neq I$

$= I \text{ (at } 720^\circ)$

Key observations:

- At $\theta = 2\pi$ (360°): the rotation $R(2\pi) = I$ has returned to the identity. But $U(2\pi) = e^{i\pi\hat{n}\cdot\boldsymbol{\sigma}} = -I \neq I$. The rotation is back; the $SU(2)$ element is not.
- At $\theta = 4\pi$ (720°): now $U(4\pi) = I$, and both the rotation and the $SU(2)$ element have returned to their starting values.

A spinor ψ that transforms as $\psi \mapsto U\psi$ (under the fundamental representation of $SU(2)$, not the adjoint) therefore acquires a sign after a 360° rotation:

$$\psi \mapsto U(2\pi)\psi = -I \cdot \psi = -\psi. \quad (25.47)$$

A vector (acting via Ad_U) is unchanged by 360° . A spinor (acting via U directly) changes sign. Only a 720° rotation is the true identity for spinors.

This is not a peculiarity of the mathematics. It is physically observable: neutrons in an interferometer have been shown to acquire a phase of -1 after a 360° rotation and to return to $+1$ after 720° —the spin-statistics theorem and the Pauli exclusion principle are rooted in this sign.

Foreshadowing the Lorentz group

The construction above is 3-dimensional, but it extends immediately to 4-dimensional Minkowski spacetime—and this extension is the foundation of relativistic quantum mechanics.

Replace the Hermitian traceless 2×2 matrix with a general Hermitian 2×2 matrix. A 4-vector $x^\mu = (x^0, x^1, x^2, x^3)$ is encoded as:

$$X = x^\mu \sigma_\mu = x^0 I + x^1 \sigma_1 + x^2 \sigma_2 + x^3 \sigma_3 = \begin{pmatrix} x^0 + x^3 & x^1 - ix^2 \\ x^1 + ix^2 & x^0 - x^3 \end{pmatrix}, \quad (25.48)$$

where $\sigma_0 = I$. Now the determinant gives the Minkowski norm:

$$\det X = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 = x^\mu x_\mu. \quad (25.49)$$

A Lorentz transformation preserves the Minkowski norm, i.e. preserves $\det X$. For the proper orthochronous Lorentz group $SO^+(1, 3)$, this can be written as:

$$X \mapsto L X L^\dagger, \quad L \in SL(2, \mathbb{C}), \quad (25.50)$$

where $\det L = 1$. This is the direct Lorentz analogue of the $SU(2)$ adjoint action—but now L need not be unitary; it is only required to have unit determinant. The group $SL(2, \mathbb{C})$ is the double cover of $SO^+(1, 3)$, with the same kernel $\{I, -I\}$.

The two-sided action $X \mapsto L X L^\dagger$ acts on the 4-vector using L on the left and $L^\dagger$ on the right—the $(j_L, j_R) = (\frac{1}{2}, \frac{1}{2})$ representation of $SL(2, \mathbb{C})$. The elementary spinors acted on by L alone (from the left, with no $L^\dagger$ on the right) are the left-handed Weyl spinors ψ_L of the Standard Model; those acted on by $(L^\dagger)^{-1}$ alone are the right-handed Weyl spinors ψ_R . The Dirac spinor is their direct sum.

Space	Encoding	Invariant	Symmetry	Double cover
$\mathbb{R}^3$	$V = \mathbf{v} \cdot \boldsymbol{\sigma}$ (Herm., traceless)	$\det V = - \mathbf{v} ^2$	$SO(3)$	$SU(2)$
$\mathbb{R}^{1,3}$	$X = x^\mu \sigma_\mu$ (Hermitian)	$\det X = x^\mu x_\mu$	$SO^+(1, 3)$	$SL(2, \mathbb{C})$

In both cases: L and $-L$ give the same transformation; spinors need 720° to return.

The 2×2 matrix encoding of 3D vectors is not a trick—it is the first step in the chain: Pauli matrices $\rightarrow$ double cover $\rightarrow$ spinors $\rightarrow$ Weyl spinors $\rightarrow$ the spinor index notation $m, \bar{m}$ of the Standard Model. Everything in this chain is already visible in the simple formula $V = \mathbf{v} \cdot \boldsymbol{\sigma}$ and the adjoint action $V \mapsto UVU^\dagger$.

25.12 The Schrödinger equation as a representation of time translation

The cleanest example of the noun/verb/equivariance picture is the one we have been using throughout this book.

Time translation is a one-parameter group: $(\mathbb{R}, +)$, with group elements $t \in \mathbb{R}$ acting on spacetime by $x^\mu \mapsto x^\mu + t\delta_0^\mu$. A representation of this group on a vector space V is a group homomorphism:

$$\rho : (\mathbb{R}, +) \rightarrow GL(V), \quad t \mapsto M(t) = e^{-iHt}, \quad (25.51)$$

for some linear operator H on V . The homomorphism property $\rho(t+s) = \rho(t)\rho(s)$ follows from $e^{-iH(t+s)} = e^{-iHt}e^{-iHs}$. The generator of this one-parameter group is $-iH$ —the Lie algebra element corresponding to an infinitesimal time translation.

The Schrödinger equation is exactly the infinitesimal version of (25.51):

$$\frac{d}{dt} e^{-iHt} h_0 = -iH e^{-iHt} h_0 \quad \Longleftrightarrow \quad \dot{h}_t = -iH h_t. \quad (25.52)$$

It is the statement that the Fock hidden state h_t transforms under time translation by the representation matrix e^{-iHt} .

Commutative diagram for time evolution.

$$\begin{array}{ccc} t & \xrightarrow{+dt} & t + dt \\ \downarrow h & & \downarrow h \\ h_t & \xrightarrow{e^{-iHdt}} & h_{t+dt} \end{array}$$

The top arrow is time translation in $\mathbb{R}$; the bottom arrow is the action of its representation matrix $e^{-iHdt} \approx I - iHdt$ on the hidden state. Commutativity says: advancing time by dt and then reading out the state gives the same result as reading out the state and then applying the evolution matrix. This is the Schrödinger equation.

H is the Lie algebra generator. The Lie algebra of $(\mathbb{R}, +)$ is $\mathbb{R}$ itself (one generator, the number 1). The Lie algebra representation is:

$$d\rho(1) = -iH, \quad (25.53)$$

so the Hamiltonian H is (up to the factor $-i$) the representation of the single generator of time translation. Noether's theorem for time translation says: if H does not depend on t (the Lagrangian is time-translation invariant), then H is conserved—energy is conserved. The conserved charge of time translation is the Hamiltonian itself.

Unitary representation and probability conservation. For the representation to preserve the norm $\|h_t\| = 1$ (probability conservation), the representation matrices must be unitary: $\rho(t)^\dagger \rho(t) = e^{iHt} e^{-iHt} = I$. This requires $H = H^\dagger$ (Hermitian). The Schrödinger equation with a Hermitian Hamiltonian is therefore the unique *unitary* representation of time translation on a Hilbert space: unitarity of the group representation forces Hermiticity of the generator, which forces real eigenvalues (real energies), which forces probability conservation. The entire probabilistic structure of quantum mechanics follows from demanding that the representation of time translation be unitary.

Energy eigenstates as irreducible representations. The irreducible representations of $(\mathbb{R}, +)$ are all one-dimensional: $\rho_E(t) = e^{-iEt}$ for $E \in \mathbb{R}$. An energy eigenstate h_E with $Hh_E = Eh_E$ transforms under time translation by the one-dimensional representation e^{-iEt} :

$$h_E(t) = e^{-iEt} h_E(0). \quad (25.54)$$

The energy E is the label of the irrep—the eigenvalue of the generator H . The general state $h_t = e^{-iHt} h_0$ decomposes into energy eigenstates:

$$h_t = \sum_E c_E e^{-iEt} h_E, \quad (25.55)$$

which is exactly the decomposition of $L^2(\mathbb{R})$ into irreps of $(\mathbb{R}, +)$ via Fourier transform: the energy eigenstates are the “spherical harmonics” of time translation, and the Fourier transform is the equivariant embedding.

Time translation	$SO(3)$ on S^2	General
Group $(\mathbb{R}, +)$	$SO(3)$	G
Generator $1 \in \mathbb{R}$	$J^3 \in \mathfrak{so}(3)$	$X \in \mathfrak{g}$
Rep. matrix e^{-iHt}	$D^{(\ell)}(R)$	$\rho(g)$
Lie algebra rep. $-iH$	$d\rho(J^3)$	$d\rho(X)$
Eigenstates h_E	Spherical harmonics Y_ℓ^m	Weight vectors
Energy E	Weight m	Weight eigenvalue
Fourier decomp. (25.55)	Spherical harmonic expansion	Peter-Weyl theorem
Unitarity $\Rightarrow H^\dagger = H$	Unitarity $\Rightarrow d\rho(X)^\dagger = -d\rho(X)$	Unitary rep.

Nouns and verbs, revisited. In the commutative diagram, $v(x)$ is the embedding of an object (a *noun*: a point, a molecule, a particle configuration). $M(g)$ is the representation of a transformation (a *verb*: a rotation, a boost, a gauge transformation). The representation $\rho : G \rightarrow GL(V)$ assigns a matrix verb $M(g)$ to every abstract verb $g \in G$. The equivariance condition (25.33) is the requirement that verbs act consistently on noun embeddings: acting on the noun and then embedding gives the same result as embedding and then applying the matrix verb.

In the Standard Model: the noun is a particle configuration $|n; \mathbf{k}_1, I_1, \dots\rangle$; the verb is a Lorentz transformation Λ or a gauge transformation $U(x)$; the representation matrices $D^{(j)}(L)$, U^α_β , V^i_j are the matrix verbs that act on the spinor and gauge indices of the field. The Schrödinger RNN update $h_t \rightarrow e^{-iHdt} h_t$ is a time-evolution verb applied to the noun h_t . The requirement that H commutes with all symmetry transformations (Noether’s theorem) is exactly the requirement that time evolution is an equivariant map with respect to the symmetry group.

26 Equivariant Representations of the Lorentz Group

26.1 Fields as equivariant representations

A field is an assignment of a mathematical object — a number, a vector, a spinor — to each point of spacetime. The object at each point is the *field variable*: it carries the local physical degrees of freedom. A scalar field $\phi(x)$ assigns a single complex number to each spacetime point; a vector field $A^\mu(x)$ assigns a four-component vector; a spinor field $\psi_A(x)$ assigns a multi-component spinor. The different types of field variable are distinguished by how they transform when we move between observers related by a Lorentz transformation. This transformation rule is the *equivariant representation*: the precise sense in which the field “knows about” the Lorentz symmetry of spacetime.

Active transformation. The most transparent way to state the transformation rule is the *active* viewpoint. A Lorentz transformation Λ acts on spacetime by moving each point x to a new point $x' = \Lambda x$. Simultaneously, the field variable at x is transformed to a new field variable at x' . The transformation of the field variable is linear in Λ and determined by the type of field.

For a scalar field:

$$\phi(x) \mapsto \phi'(x') = \phi(x), \quad (26.1)$$

the field variable at the new point $x' = \Lambda x$ is the same number as at the original point x : the scalar does not rotate.

For a vector field:

$$A^\mu(x) \mapsto A'^\mu(x') = \Lambda^\mu{}_\nu A^\nu(x), \quad (26.2)$$

the field variable at $x' = \Lambda x$ is obtained by applying the matrix Λ to the components of $A^\mu(x)$: the vector rotates with the spacetime point.

For a spinor field:

$$\psi_A(x) \mapsto \psi'_A(x') = S(\Lambda)^A{}_B \psi_B(x), \quad (26.3)$$

the field variable at $x' = \Lambda x$ is obtained by applying a matrix $S(\Lambda)$ to the spinor components. The matrix $S(\Lambda)$ is not the same as Λ itself: it is a different representation of the same Lorentz transformation, acting on a different vector space (the spinor space rather than the spacetime tangent space). Which matrix $S(\Lambda)$ is used depends on the *representation* of the Lorentz group assigned to the field.

In all three cases, the rule has the same structure: the spacetime point transforms by Λ , and the field variable at that point transforms by a linear map $\rho(\Lambda)$ (which is 1 for scalars, Λ for vectors, and $S(\Lambda)$ for spinors). Written uniformly:

$$\Phi(x) \mapsto \Phi'(\Lambda x) = \rho(\Lambda) \Phi(x). \quad (26.4)$$

This is the definition of an *equivariant map* in the language of representation theory: the field Φ intertwines the action of Λ on spacetime (moving the point x to Λx) with the action of $\rho(\Lambda)$ on the field variable (rotating the internal components). The two actions are consistent: if we apply two successive Lorentz transformations Λ_2 and Λ_1 , the field transforms by $\rho(\Lambda_2)\rho(\Lambda_1) = \rho(\Lambda_2\Lambda_1)$, which is the same as applying the composed transformation $\Lambda_2\Lambda_1$ directly. This is the representation property: ρ is a group homomorphism from the Lorentz group to the group of linear transformations of the field variable.

Connection to equivariant representations in learning. This structure is exactly the *equivariant representation* familiar from geometric deep learning. In an equivariant neural network, a feature map $\Phi(x)$ at position x transforms consistently when a symmetry transformation g is applied: the feature at the transformed position gx equals the representation $\rho(g)$ applied

to the feature at x . The Standard Model fields are equivariant feature maps on spacetime, with the Lorentz group playing the role of the symmetry group, and the representation $\rho(\Lambda)$ (scalar, vector, spinor, etc.) playing the role of the feature transformation. A scalar field is a trivial equivariant map (the feature does not transform); a vector field is equivariant under the defining representation; a spinor field is equivariant under the spinor representation. The study of which representations $\rho(\Lambda)$ are possible is precisely the representation theory of the Lorentz group, and it determines the complete list of field types that can appear in a Lorentz-invariant theory.

Passive transformation. The same physics can be described equivalently by the *passive* viewpoint: instead of moving the spacetime point x to $x' = \Lambda x$, we hold all spacetime points fixed and transform the coordinate system by Λ^{-1} . In the new coordinate system, the point that was labelled x' is now labelled x , and the field variable that was $\Phi'(x')$ in the new coordinates is the same physical object as $\Phi(x)$ in the old coordinates. The transformation rule (26.4) in the active language becomes, in the passive language:

$$\Phi_{\text{new}}(x) = \rho(\Lambda) \Phi_{\text{old}}(\Lambda^{-1}x), \quad (26.5)$$

which says: the field in the new coordinate system at point x equals the representation $\rho(\Lambda)$ applied to the field in the old coordinate system at the point $\Lambda^{-1}x$ (the old label for the same physical spacetime point). Both viewpoints are correct; the active viewpoint is often more intuitive for understanding what the transformation does, while the passive viewpoint is more natural for deriving transformation rules from first principles.

From equivariant fields to a Lorentz-invariant Lagrangian. Once the fields $\Phi(x)$ are assigned equivariant representations, building a Lorentz-invariant Lagrangian reduces to a problem in linear algebra: contract all Lorentz indices using the Minkowski metric $\eta^{\mu\nu}$ and the invariant tensors of the Lorentz group (such as the spinor metric ϵ_{AB} and the gamma matrices γ_{AB}^μ) so that the result is a scalar under (26.4). For example, the scalar kinetic term $\partial_\mu \phi \partial^\mu \phi = \eta^{\mu\nu} \partial_\mu \phi \partial_\nu \phi$ is a Lorentz scalar built from two vector-like objects (the gradient $\partial_\mu \phi$ transforms as a covariant vector under Λ), contracted with the inverse metric $\eta^{\mu\nu}$. The Dirac kinetic term $\bar{\psi} \gamma^\mu \partial_\mu \psi = \bar{\psi}^A (\gamma^\mu)_{AB} \partial_\mu \psi^B$ is a scalar built by contracting spinor indices with the gamma matrices and the spacetime index with ∂_μ .

Once a scalar Lagrangian $\mathcal{L}[\Phi, \partial_\mu \Phi]$ is in hand, the remainder of the construction follows the standard route:

$$\underbrace{\mathcal{L}[\Phi, \partial_\mu \Phi]}_{\text{Lorentz scalar}} \xrightarrow{\text{Legendre}} \underbrace{H[\Phi, \pi]}_{\text{Hamiltonian}} \xrightarrow{\text{quantize}} \underbrace{H[a^\dagger, a]}_{\text{Fock operator}} \xrightarrow{\text{time slice}} \underbrace{\int \mathcal{D}\Phi e^{iS}}_{\text{path integral}} \xrightarrow{\text{Gaussian integral}} \mathcal{L}. \quad (26.6)$$

The equivariance of the fields guarantees that the Lorentz scalar property of $\mathcal{L}$ is preserved at every step of this round trip. The Fock basis vectors $|n; \mathbf{k}_j, s_j, \dots\rangle$ inherit their spinor and Lorentz indices from the equivariant representations of the fields; the creation and annihilation operators $a^\dagger(\mathbf{k}, s)$ carry the same indices; and the Born rendering layer projects onto physical states that transform in definite representations of the Lorentz group.

The task of the present section is to classify all equivariant representations $\rho(\Lambda)$ of the Lorentz group: the complete list of field types that can appear in (26.4). This classification is the representation theory of the Lorentz group.

26.2 Diagram 1: the abstract transformation law

A field $\Phi(x)$ lives at each spacetime point x and takes values in some vector space V . Under a Lorentz transformation Λ , the point and the value transform together. This is the fundamental diagram:

$$\begin{array}{ccc}
x & \xrightarrow{\Lambda} & x' = \Lambda x \\
\downarrow & & \downarrow \\
\Phi(x) & \xrightarrow{M(\Lambda)} & \Phi(x') = M(\Lambda) \Phi(x)
\end{array}$$

The vertical arrows are evaluation of the field at a point. The top arrow is the Lorentz transformation on spacetime. The bottom arrow is the induced action on the field value. The diagram commutes: transform the point and then evaluate, or evaluate and then transform the value—you get the same answer.

The assignment $\Lambda \mapsto M(\Lambda)$ must be a group homomorphism: $M(\Lambda_1 \Lambda_2) = M(\Lambda_1) M(\Lambda_2)$. This is the *representation condition*. Different choices of M give different species of field.

ML reading. The field $\Phi(x)$ is an equivariant embedding table: each row is the “embedding” of spacetime point x in V . A Lorentz transformation acts on the row index via Λ and on the row content via $M(\Lambda)$ simultaneously. Classifying all possible fields is classifying all equivariant representations of the Lorentz group—the same problem as finding all equivariant maps in geometric deep learning.

26.3 Diagram 2: spacetime as 2×2 matrices

To make the representation $M(\Lambda)$ explicit, we first rewrite the spacetime point $x^\mu = (x^0, x^1, x^2, x^3)$ as a 2×2 Hermitian matrix by contracting with the Pauli matrices:

$$X = x^\mu \sigma_\mu = x^0 I + x^1 \sigma_1 + x^2 \sigma_2 + x^3 \sigma_3 = \begin{pmatrix} x^0 + x^3 & x^1 - ix^2 \\ x^1 + ix^2 & x^0 - x^3 \end{pmatrix}. \quad (26.7)$$

Two facts make this the right thing to do. First, every 2×2 Hermitian matrix arises this way: the map $x^\mu \mapsto X$ is a vector-space isomorphism between $\mathbb{R}^4$ and the space of 2×2 Hermitian matrices. Second, the determinant of X is the spacetime interval:

$$\det X = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2 = x^\mu x_\mu. \quad (26.8)$$

Now let $L \in SL(2, \mathbb{C})$ (a complex 2×2 matrix with $\det L = 1$) act on X by *conjugation*:

$$X' = L X L^\dagger. \quad (26.9)$$

Then X' is still Hermitian ($X'^\dagger = (L X L^\dagger)^\dagger = L X L^\dagger = X'$), and its determinant is preserved: $\det X' = \det L \cdot \det X \cdot \det L^\dagger = \det X$, since $\det L = 1$ and $\det L^\dagger = \overline{\det L} = 1$. Therefore X' corresponds to a new spacetime point x'^μ with the same spacetime interval as x^μ : the map $X \mapsto L X L^\dagger$ is a Lorentz transformation $\Lambda(L)$.

$$\begin{array}{ccc}
x & \xrightarrow{\Lambda} & x' \\
\downarrow & & \downarrow \\
X = x^\mu \sigma_\mu & \xrightarrow{X' = L X L^\dagger} & X' = x'^\mu \sigma_\mu
\end{array}$$

The map $L \mapsto \Lambda(L)$ is a 2-to-1 group homomorphism $SL(2, \mathbb{C}) \rightarrow SO^+(3, 1)$: the two matrices L and $-L$ give the same Λ . This is why the Lorentz group has “double covers” and why half-integer spin is possible.

L and $L^\dagger$ together represent Λ . The matrix L acts on the left index of X and $L^\dagger$ acts on the right index. These are the two independent $SL(2, \mathbb{C})$ factors that arise from the factorisation of the Lorentz algebra $\mathfrak{so}(3, 1)_\mathbb{C} \cong \mathfrak{sl}(2, \mathbb{C}) \oplus \mathfrak{sl}(2, \mathbb{C})$, labelled (j_L, j_R) in the previous section. For real spacetime vectors, L and $L^\dagger$ are conjugates of each other (not independent), so they jointly encode the single Lorentz transformation Λ . Different choices of how to split the representation between left and right give the different particle types.

Example. Rotation and boost in $SL(2, \mathbb{C})$.

For a rotation by θ about $\hat{z}$:

$$L = e^{-i\theta\sigma_3/2} = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}, \quad L^\dagger = \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix}.$$

One can verify that $X' = LXL^\dagger$ rotates (x^1, x^2) by angle θ while leaving x^0 and x^3 unchanged—exactly the Lorentz rotation Λ .

For a boost with rapidity ξ along $\hat{z}$:

$$L = e^{-\xi\sigma_3/2} = \begin{pmatrix} e^{-\xi/2} & 0 \\ 0 & e^{\xi/2} \end{pmatrix}, \quad L^\dagger = \begin{pmatrix} e^{-\xi/2} & 0 \\ 0 & e^{\xi/2} \end{pmatrix}.$$

(Both real here since σ_3 is real.) $X' = LXL^\dagger$ mixes x^0 and x^3 as a hyperbolic rotation—exactly the Lorentz boost.

26.4 Diagram 3: fields on matrix spacetime

Now we promote X from a label for a spacetime point to the *argument* of the field. The field value $\Phi(X)$ is itself a matrix, and the Lorentz group acts on both its left and right indices:

$$\Phi(X) \mapsto \Phi(X') = M_L(L) \Phi(X) M_R(L^\dagger). \quad (26.10)$$

$$\begin{array}{ccc} X & \xrightarrow{X'=LXL^\dagger} & X' \\ \downarrow & & \downarrow \\ \Phi(X) & \xrightarrow{\Phi(X')=M_L(L)\Phi(X)M_R(L^\dagger)} & \Phi(X') \end{array}$$

This is the master transformation law. Every particle species corresponds to a specific choice of the left representation M_L and the right representation M_R :

Field	$\Phi(X)$ shape	$M_L(L)$	$M_R(L^\dagger)$	Particle
Scalar	1×1 (number)	1	1	Higgs
Left Weyl	2×1 (column)	L	1	ν_L, e_L
Right Weyl	2×1 (column)	1	$(L^\dagger)^T$	e_R
4-vector	2×2 (matrix)	L	$L^\dagger$	$\gamma, W^\pm, Z$

Remark. [Two meanings of “vector”] In this table, the $(\frac{1}{2}, \frac{1}{2})$ representation is called a *4-vector* or simply a *vector* in the physics sense: it is an object with four spacetime components that rotates and boosts like x^μ . This is a representation-theoretic statement about how the object transforms. It has nothing to do with the shape of the array storing the components.

By contrast, when we say that the left Weyl spinor is a 2×1 column vector and the Dirac spinor is a 4×1 column vector, we mean *column vector in the linear-algebra sense*: a one-column array of complex numbers on which matrices act by left multiplication. Please keep these two usages distinct; the collision of vocabulary is unfortunately standard in physics.

Reading across the rows:

Scalar. $M_L = M_R = 1$; the field value is a single number, blind to both the left and right actions of L . Diagram 3 reduces to Diagram 1 with $M(\Lambda) = 1$.

Left Weyl spinor. $M_L = L$, $M_R = 1$; the field value is a 2×1 column vector and only the left $SL(2, \mathbb{C})$ factor acts:

$$\psi_L(X) \mapsto L \psi_L(X). \quad (26.11)$$

A 2×2 matrix L multiplies the 2×1 column on the *left*—the name “left Weyl” is precisely this. This is the $(\frac{1}{2}, 0)$ representation.

Right Weyl spinor. The right Weyl spinor requires more care about index placement, and getting it right pays dividends when we build the Dirac spinor.

The natural home of ψ_R in Diagram 3 is as a *row vector*: the right $SL(2, \mathbb{C})$ factor acts on the right index of the 2×2 matrix spacetime X , and a field carrying only a right index is naturally a 1×2 row. Diagram 3 with $M_L = 1$, $M_R = L^\dagger$ gives

$$\psi_R(X)^T \mapsto \psi_R(X)^T L^\dagger, \quad \psi_R(X)^T \in \mathbb{C}^{1 \times 2}. \quad (26.12)$$

The 1×2 row is multiplied by the 2×2 matrix $L^\dagger$ on the *right*—hence “right Weyl”.

However, for everything that follows (stacking spinors, writing Lagrangians, acting with Dirac matrices) we want *all* spinors to be 2×1 columns. We therefore transpose equation (26.12). Using $(AB)^T = B^T A^T$:

$$(\psi_R^T)^T \mapsto (\psi_R^T L^\dagger)^T = (L^\dagger)^T \psi_R. \quad (26.13)$$

Now $(L^\dagger)^T$: recall $L^\dagger = (\bar{L})^T$, so $(L^\dagger)^T = \bar{L}$, the *complex conjugate* of L (transpose the transpose, conjugate stays). Denoting the 2×1 column simply as ψ_R , the transformation law in column form is

$$\psi_R(X) \mapsto \bar{L} \psi_R(X), \quad \psi_R(X) \in \mathbb{C}^{2 \times 1}. \quad (26.14)$$

	Shape	Transformation
Left Weyl	2×1 column	$\psi_L \mapsto L \psi_L$
Right Weyl	2×1 column	$\psi_R \mapsto \bar{L} \psi_R$

Both spinors are now 2×1 columns acted on by 2×2 matrices on the left. The difference is entirely in *which* 2×2 matrix acts: the original L for the left Weyl spinor, and the complex conjugate $\bar{L}$ for the right Weyl spinor.

4-vector. $M_L = L$, $M_R = L^\dagger$; the field value is a 2×2 matrix and both factors act:

$$A(X) \mapsto L A(X) L^\dagger. \quad (26.15)$$

This is the $(\frac{1}{2}, \frac{1}{2})$ representation. The field value transforms *in exactly the same way as the spacetime point X itself*: the bottom arrow in Diagram 3 is identical to the bottom arrow in Diagram 2. The photon field A_μ , written as $A = A_\mu \sigma^\mu$, is a second copy of spacetime living fibre-wise over spacetime.

26.5 Chirality

The word *chirality* refers to the handedness of a Weyl spinor—which of the two $SL(2, \mathbb{C})$ factors in the Lorentz algebra acts on it. Here is the mnemonic that makes it impossible to confuse the two:

Left-chiral ψ_L	Right-chiral ψ_R
2×1 column vector	2×1 column vector
Acted on by matrix L on the left	Acted on by matrix $\bar{L}$ on the left (or equivalently $L^\dagger$ on the right before transposing)
$(\frac{1}{2}, 0)$ representation	$(0, \frac{1}{2})$ representation
Carries a <i>left</i> $SL(2, \mathbb{C})$ index	Carries a <i>right</i> $SL(2, \mathbb{C})$ index
Neutrinos ν_L , left-handed electron e_L	Right-handed electron e_R

The mnemonic: *a left Weyl spinor is a column acted on by a matrix on the left*. A right Weyl spinor also ends up as a column acted on by a matrix on the left—but that matrix is $\bar{L}$, not L , a reminder that it originally lived as a row acted on from the right.

Why two chiralities? The Lorentz algebra over $\mathbb{C}$ splits as $\mathfrak{so}(3, 1)_\mathbb{C} \cong \mathfrak{sl}(2, \mathbb{C}) \oplus \mathfrak{sl}(2, \mathbb{C})$, and any representation is labelled by how it transforms under each factor. The “left” factor generated by $\mathbf{J} - i\mathbf{K}$ acts on left-chiral spinors; the “right” factor generated by $\mathbf{J} + i\mathbf{K}$ acts on right-chiral spinors. (Here $\mathbf{J}$ are the angular momentum generators and $\mathbf{K}$ the boost generators.)

26.6 Spinors versus vectors: the sign that makes all the difference

The difference between a spinor and a vector is not merely a matter of the number of components. It is a statement about *how many times the Lorentz matrix L appears* in the transformation law.

Compare the two transformation laws side by side:

$$\psi_L(X) \mapsto L \psi_L(X), \quad (26.16)$$

$$A(X) \mapsto L A(X) L^\dagger. \quad (26.17)$$

The spinor sees *one* copy of L . The vector sees L on the left *and* $L^\dagger$ on the right—two insertions of the group element. This counting is everything.

Tracking a full 2π rotation. A rotation by angle θ about the z -axis is represented in $SL(2, \mathbb{C})$ by

$$L(\theta) = e^{-i\frac{\theta}{2}\sigma_z} = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}. \quad (26.18)$$

At $\theta = 0$ we have $L = +I$. As θ increases from 0 to 2π —a physically complete rotation—the matrix L travels continuously from $+I$ to $-I$. Substituting these two endpoints into each transformation law:

Object	Law	$L = +I$	$L = -I$
Spinor ψ_L	$\psi_L \mapsto L \psi_L$	$+\psi_L$	$-\psi_L$
Vector A	$A \mapsto L A L^\dagger$	$+A$	$(-I) A (-I)^\dagger = +A$

The vector is *oblivious* to the sign flip: the two insertions of L each contribute a factor of (-1) , and $(-1)(-1) = +1$. The spinor, carrying only one insertion, acquires the naked factor of (-1) .

A 2π rotation does not return the spinor to itself. Only a 4π rotation does, because $L(4\pi) = +I$ again.

The spinor as a “square root” of the vector The vector representation is built from L and $L^\dagger$ simultaneously; the spinor carries only L . In this precise sense a spinor is a *square root* of a vector: two spinor indices—one transforming with L and one with $L^\dagger$ —combine to produce a vector index. This is made explicit by the outer product

$$A^{s\dot{s}} = \psi_L^s \psi_R^{\dot{s}}, \quad (26.19)$$

where s is a left $SL(2, \mathbb{C})$ index and $\dot{s}$ is a right one. Multiplying the transformation laws, $L \cdot L^\dagger$ acting on the pair, recovers exactly (26.17). A single undotted index is a spinor; a dotted-undotted pair is a vector.

An analogy with ordinary complex numbers. Let the unit circle in $\mathbb{C}$ stand in for $SU(2)$, and let $z = e^{i\varphi}$ represent the group element at angle φ . A “vector-like” object transforms as $z \mapsto z \cdot \bar{z} = |z|^2 = 1$; the phase cancels. A “spinor-like” object transforms as $z \mapsto z$; the phase survives. Winding φ from 0 to 2π takes z from $+1$ to $+1$ (two factors cancel), but takes a single z from $+1$ to $+1$ *only after* winding to 4π —one factor, one full trip around the circle. The Lorentz story is the non-abelian version of exactly this.

Physical consequence: interference, not just bookkeeping. The minus sign acquired by a spinor under 2π rotation is *unobservable in isolation*—a global phase never appears in any expectation value. It becomes observable only when two paths in parameter space interfere. The classic demonstration is the neutron interferometry experiment of Werner *et al.* (1975): a neutron beam is split, one arm is subjected to a magnetic field that rotates the spin, and the two arms are recombined. The interference pattern shifts by exactly π when the spin rotation passes through 2π , confirming that the spinor picks up -1 . The vector (photon, classical wave) would show no such shift. This is a direct physical measurement of the two-to-one covering $SU(2) \rightarrow SO(3)$.

26.7 General representations: (j_L, j_R) as matrix dimensions

The pattern is now clear. For a general (j_L, j_R) representation, the field value $\Phi(X)$ lives in a $(2j_L + 1) \times (2j_R + 1)$ matrix, and:

$$\Phi(X) \mapsto D^{(j_L)}(L) \Phi(X) D^{(j_R)}(L^\dagger), \quad (26.20)$$

where $D^{(j)}(L)$ is the $(2j + 1)$ -dimensional spin- j representation of $SL(2, \mathbb{C})$. The left and right matrix dimensions count the number of left-handed and right-handed spinor indices. Scalars have no indices (1×1); Weyl spinors have one index (2×1 or 1×2); 4-vectors have one index on each side (2×2); higher-spin particles have more.

26.8 Dirac spinor

Building the Dirac spinor by direct sum. We now have two 2×1 column vectors—a left-chiral spinor ψ_L and a right-chiral spinor ψ_R —each living in its own $\mathbb{C}^2$. The *direct sum* $\mathbb{C}^2 \oplus \mathbb{C}^2$ means, concretely, that we *concatenate* the two columns into a single 4×1 column:

$$\Psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} = \begin{pmatrix} \psi_{L,1} \\ \psi_{L,2} \\ \psi_{R,1} \\ \psi_{R,2} \end{pmatrix} \in \mathbb{C}^4. \quad (26.21)$$

This is the *Dirac spinor*. The upper two components are left-chiral; the lower two are right-chiral. No mixing has occurred yet—the direct sum is just bookkeeping, placing two independent objects side by side in one column.

Because Ψ is a 4×1 column with no free right index, any linear transformation of Ψ must be a 4×4 matrix acting on the left. Using the column-vector transformation laws (26.11) and (26.14), the upper block transforms by L and the lower block by $\bar{L}$, so the 4×4 matrix is block-diagonal:

$$\Psi \mapsto S(L) \Psi, \quad S(L) = \begin{pmatrix} L & 0 \\ 0 & \bar{L} \end{pmatrix}. \quad (26.22)$$

Remark. [Why $\bar{L}$, not $L^\dagger$?] In the row-vector form (26.12), the right Weyl spinor transformed as $\psi_R^T \mapsto \psi_R^T L^\dagger$. After transposing to a column vector we found $\psi_R \mapsto \bar{L} \psi_R$. The Dirac matrix $S(L)$ must act on *column* vectors, so the correct entry is $\bar{L}$, the complex conjugate of L . Writing $L^\dagger$ would be a shape error: $L^\dagger$ appeared when ψ_R was a row being right-multiplied; once we transpose to a column, the matrix that appears on the left is $(L^\dagger)^T = \bar{L}$.

This is the $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation: a direct sum of the left and right Weyl representations. The two chiralities do not talk to each other—yet.

Why bother with the Dirac spinor? Enter the gamma matrices.

If the Dirac spinor were merely a notational convenience for carrying two Weyl spinors around together, it would not be worth introducing. The payoff is that the 4×4 algebra of Dirac matrices—the *gamma matrices*—acts on Ψ in a way that *mixes* ψ_L and ψ_R , and this mixing is precisely what a mass term requires.

The gamma matrices γ^μ ($\mu = 0, 1, 2, 3$) are four 4×4 complex matrices satisfying the *Clifford algebra*:

$$\{\gamma^\mu, \gamma^\nu\} \equiv \gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2\eta^{\mu\nu} \mathbf{1}_{4 \times 4}, \quad (26.23)$$

where $\eta^{\mu\nu} = \text{diag}(+1, -1, -1, -1)$ is the Minkowski metric. In the *Weyl basis* (also called chiral basis), which is the natural basis for our block decomposition, we first define four 2×2 matrices by enumerating $\mu = 0, 1, 2, 3$:

$$\sigma^\mu = (I, \sigma^1, \sigma^2, \sigma^3) = (I, \vec{\sigma}), \quad \bar{\sigma}^\mu = (I, -\sigma^1, -\sigma^2, -\sigma^3) = (I, -\vec{\sigma}), \quad (26.24)$$

where I is the 2×2 identity and $\vec{\sigma} = (\sigma^1, \sigma^2, \sigma^3)$ are the Pauli matrices. Then, for each fixed μ , the corresponding gamma matrix is the 4×4 matrix assembled from these 2×2 blocks:

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{pmatrix}. \quad (26.25)$$

Each γ^μ is block *off*-diagonal: it maps the upper (ψ_L) block to the lower (ψ_R) block and vice versa. One can verify directly that the Clifford algebra (26.23) is satisfied using the Pauli identity $\sigma^i \sigma^j = \delta^{ij} I + i\epsilon^{ijk} \sigma^k$.

Now consider the *Dirac equation*:

$$(i\gamma^\mu \partial_\mu - m)\Psi = 0. \quad (26.26)$$

Writing this out in blocks:

$$\begin{pmatrix} -m & i\sigma^\mu \partial_\mu \\ i\bar{\sigma}^\mu \partial_\mu & -m \end{pmatrix} \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} = 0, \quad (26.27)$$

which gives two coupled equations:

$$i \sigma^\mu \partial_\mu \psi_R = m \psi_L, \quad (26.28)$$

$$i \bar{\sigma}^\mu \partial_\mu \psi_L = m \psi_R. \quad (26.29)$$

Three features are immediately visible:

1. **Mass couples the two chiralities.** If $m = 0$, equations (26.28)–(26.29) decouple completely: ψ_L and ψ_R are independent. The photon and the neutrino (in the massless limit) are described by a single Weyl spinor. A nonzero mass m forces the two chiralities to communicate: knowing ψ_R constrains ψ_L , and vice versa. The Dirac spinor is therefore not a luxury—it is the *minimal* object that can describe a massive spin- $\frac{1}{2}$ particle.
2. **Gamma matrices are the off-diagonal mixing operator.** The γ^μ in equation (26.26) are 4×4 matrices because they must act on the full 4×1 Dirac spinor. Their off-diagonal block structure in (26.25) is what implements the $\psi_L \leftrightarrow \psi_R$ coupling. The Clifford algebra (26.23) encodes the Minkowski metric and ensures that the Dirac equation is Lorentz-covariant.
3. **ML analogy.** Think of ψ_L and ψ_R as two separate feature channels of a neural network layer. In the massless limit they are processed by separate pathways. The mass term m introduces *cross-channel mixing*, exactly as an off-diagonal weight matrix would mix two feature groups. The Dirac equation is then the constraint that the “output” of each channel, after mixing, must vanish—the on-shell condition for a free massive fermion.

Summary. The hierarchy is:

$$\underbrace{\psi_L, \psi_R}_{\text{Weyl spinors: } 2 \times 1 \text{ columns}} \xrightarrow{\text{concatenate}} \underbrace{\Psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix}}_{\text{Dirac spinor: } 4 \times 1 \text{ column}} \xrightarrow{\gamma^\mu \text{ acts}} \underbrace{\text{mixing of chiralities}}_{\text{mass, interactions}}.$$

The Pauli matrices σ^μ (already familiar) reappear as the off-diagonal blocks of the gamma matrices, which is why the 2×2 language of $SL(2, \mathbb{C})$ and the 4×4 language of the Dirac equation are two faces of the same structure.

The Dirac conjugate and Lorentz-invariant bilinears

Having assembled the Dirac spinor $\Psi = (\psi_L, \psi_R)^T \in \mathbb{C}^4$, we immediately face a problem: how do we form a Lorentz-invariant scalar from Ψ ?

The naive attempt fails. The most natural candidate is $\Psi^\dagger \Psi$:

$$\Psi^\dagger \Psi = \psi_L^\dagger \psi_L + \psi_R^\dagger \psi_R. \quad (26.30)$$

Under a Lorentz transformation with $S(L) = \text{diag}(L, \bar{L})$:

$$\Psi \mapsto S(L)\Psi, \quad \Psi^\dagger \mapsto \Psi^\dagger S(L)^\dagger = \Psi^\dagger \begin{pmatrix} L^\dagger & 0 \\ 0 & \bar{L}^\dagger \end{pmatrix}. \quad (26.31)$$

For the product to be invariant we would need $S(L)^\dagger S(L) = I$, i.e. $S(L)$ unitary. But $S(L)$ is *not* unitary for boosts: under a boost with rapidity ξ , $L = e^{-\xi \sigma_3/2}$ is real and symmetric, so $L^\dagger = L \neq L^{-1}$. The combination $\Psi^\dagger \Psi$ is therefore not Lorentz invariant.

The correct invariant: mix the two chiralities. The Lorentz transformation acts as L on ψ_L and as $\bar{L}$ on ψ_R . Note that L and $\bar{L}$ are related by:

$$(L^\dagger)^{-1} = \bar{L} \quad \text{for } L \in SL(2, \mathbb{C}), \quad (26.32)$$

which follows from $\det L = 1$. This means $\psi_L^\dagger$ transforms as $(L^\dagger)^{-1} = \bar{L}$: the same as ψ_R . The cross combination:

$$\psi_L^\dagger \psi_R \mapsto \psi_L^\dagger (L^\dagger)^{-1} \bar{L} \psi_R = \psi_L^\dagger \bar{L} \bar{L}^{-1} \bar{L} \psi_R, \quad (26.33)$$

which still does not simplify cleanly. The correct pairing uses the other cross term. Under a Lorentz transformation:

$$\psi_R^\dagger \psi_L \mapsto \psi_R^\dagger (\bar{L})^\dagger L \psi_L = \psi_R^\dagger (L^T)^{-1} L \psi_L. \quad (26.34)$$

Again this fails for general L . The resolution is to use γ^0 .

The Dirac conjugate. In the Weyl (chiral) basis, the gamma matrix γ^0 is:

$$\gamma^0 = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}, \quad (26.35)$$

which *swaps* the upper and lower blocks. Define the *Dirac conjugate*:

$$\bar{\Psi} \equiv \Psi^\dagger \gamma^0 = (\psi_L^\dagger, \psi_R^\dagger) \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} = (\psi_R^\dagger, \psi_L^\dagger). \quad (26.36)$$

The Dirac conjugate swaps the roles of the left and right components: it takes $\psi_L^\dagger$ and places it in the right slot, and $\psi_R^\dagger$ in the left slot.

Why $\bar{\Psi}\Psi$ is Lorentz invariant. Now compute $\bar{\Psi}\Psi$:

$$\bar{\Psi}\Psi = (\psi_R^\dagger, \psi_L^\dagger) \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} = \psi_R^\dagger \psi_L + \psi_L^\dagger \psi_R. \quad (26.37)$$

Under a Lorentz transformation, $\psi_L \rightarrow L\psi_L$ and $\psi_R \rightarrow \bar{L}\psi_R$, so:

$$\psi_R^\dagger \psi_L \mapsto \psi_R^\dagger \bar{L}^\dagger L \psi_L = \psi_R^\dagger \psi_L, \quad (26.38)$$

where the last equality uses $\bar{L}^\dagger L = (L^\dagger)^T L = I$ for $L \in SL(2, \mathbb{C})$ (since $\bar{L}^\dagger = (L^*)^{TT} = L^* = \bar{L}^*$... more directly: for $SL(2, \mathbb{C})$, $(L^\dagger)^{-1} = \bar{L}$ implies $\bar{L}^\dagger = L^{-1}$, so $\bar{L}^\dagger L = I$). Similarly $\psi_L^\dagger \psi_R$ is invariant. Therefore:

$$\bar{\Psi}\Psi \mapsto \bar{\Psi}\Psi. \quad \checkmark \quad (26.39)$$

The combination $\bar{\Psi}\Psi$ is a genuine Lorentz scalar. It is the *Dirac mass term* $m\bar{\Psi}\Psi = m(\psi_R^\dagger \psi_L + \psi_L^\dagger \psi_R)$: it couples the two chiralities and gives the Dirac fermion a mass.

Why γ^0 is the right matrix. The γ^0 in the Weyl basis swaps $\psi_L \leftrightarrow \psi_R$ precisely because the two chiralities transform under $(L, \bar{L})$ which are related by $(L^\dagger)^{-1} = \bar{L}$. The γ^0 insertion corrects for the mismatch between the transformation laws of $\psi_L^\dagger$ and ψ_L : without it, $\Psi^\dagger \Psi$ mixes $(j_L, j_R) = (\frac{1}{2}, 0)$ with its own conjugate, giving something that transforms non-trivially; with it, the cross pairing $\psi_R^\dagger \psi_L$ uses one field from each representation and contracts to a $(0, 0)$ scalar.

More generally, any bilinear $\bar{\Psi} \Gamma \Psi$ where Γ is a product of gamma matrices has well-defined Lorentz transformation properties determined by the index structure of Γ :

Bilinear	Lorentz type	Use
$\bar{\Psi}\Psi$	Scalar $(0, 0)$	Mass term
$\bar{\Psi}\gamma^\mu\Psi$	Vector $(\frac{1}{2}, \frac{1}{2})$	Current j^μ
$\bar{\Psi}\gamma^{\mu\nu}\Psi$	Antisym. tensor	Magnetic moment
$\bar{\Psi}\gamma^\mu\gamma^5\Psi$	Axial vector	Weak current
$\bar{\Psi}\gamma^5\Psi$	Pseudoscalar $(0, 0)$	CP violation

In every case, the Dirac conjugate $\bar{\Psi} = \Psi^\dagger \gamma^0$ is what makes the bilinear transform correctly. Without it, none of the entries in the table would have definite Lorentz transformation properties.

Feynman slash notation. The slash notation $\not{p}$ is a shorthand for the contraction of a four-vector with the gamma matrices:

$$\not{p} \equiv \gamma^\mu p_\mu = \gamma^0 p_0 - \gamma^1 p_1 - \gamma^2 p_2 - \gamma^3 p_3. \quad (26.40)$$

The result is a 4×4 matrix in Dirac space, not a number. The notation is used throughout to compress expressions involving the Dirac operator: the free Dirac equation $(i\gamma^\mu \partial_\mu - m)\Psi = 0$ is written $(i\not{\partial} - m)\Psi = 0$, the fermion propagator numerator $\gamma^\mu p_\mu + m$ is written $\not{p} + m$, and the quark-gluon vertex $\gamma^\mu A_\mu$ is written $\not{A}$. The key algebraic property of the slash follows directly from the Clifford algebra:

$$\not{p}\not{q} = \gamma^\mu \gamma^\nu p_\mu q_\nu = p \cdot q - i\sigma^{\mu\nu} p_\mu q_\nu, \quad \not{p}^2 = p^\mu p_\mu = p^2, \quad (26.41)$$

where $\sigma^{\mu\nu} = \frac{i}{2}[\gamma^\mu, \gamma^\nu]$. In particular, for an on-shell fermion with $p^2 = m^2$: $\not{p}\not{p} = m^2$.

26.9 Diagram 3 in spin-index notation

Diagram 3 tells us that the field $\Phi(X)$ transforms as $\Phi(X) \rightarrow M_L(L) \Phi(X) M_R(L^\dagger)$, with M_L acting on rows and M_R acting on columns. To work with individual components—as is necessary when writing Lagrangians, Feynman rules, and equations of motion—we must give names to these row and column indices. The result is *van der Waerden notation*, the standard index language of spinor algebra in QFT.

Undotted and dotted indices. There are two independent $SL(2, \mathbb{C})$ factors, and we need to be able to tell at a glance which factor acts on which index. The convention is:

- An *undotted* index $s \in \{1, 2\}$ (or more generally $s \in \{1, \dots, 2j_L + 1\}$) labels the **left** spinor space. It transforms under $L \in SL(2, \mathbb{C})$.
- A *dotted* index $\dot{s} \in \{1, 2\}$ (or $\dot{s} \in \{1, \dots, 2j_R + 1\}$) labels the **right** spinor space. It transforms under $\bar{L}$, the complex conjugate of L .

The dot is not decoration—it is a *type tag*: an undotted index and a dotted index live in different representation spaces and can never be contracted with each other. Whenever you see a dot, you know the index belongs to the $(0, \frac{1}{2})$ world; whenever you see no dot, it belongs to the $(\frac{1}{2}, 0)$ world.

Why $\bar{L}$ for dotted indices? Recall that after transposing the right Weyl spinor from a row to a column, its transformation law became $\psi_R \mapsto \bar{L} \psi_R$. The dotted index $\dot{s}$ is the component index of ψ_R , so it must transform with $\bar{L}$. More precisely, if ψ_R has components $\psi_R^{\dot{s}}$, then:

$$\psi_R^{\dot{s}} \mapsto \bar{L}^{\dot{s} \dot{s}'} \psi_R^{\dot{s}'}, \quad (26.42)$$

where $\bar{L}^{\dot{s} \dot{s}'} \equiv \overline{L^{s s'}}$ is the element-wise complex conjugate of L . The dot on the index is a reminder that conjugation has occurred.

Upper and lower index placement. Within each spinor space there is a further distinction between upper and lower indices, analogous to contravariant and covariant indices in tensor calculus. The rule is:

- *Upper* index: transforms with L (undotted) or $\bar{L}$ (dotted).
- *Lower* index: transforms with $(L^{-1})^T$ (undotted) or $(\bar{L}^{-1})^T$ (dotted).

For $SL(2, \mathbb{C})$ the two-index antisymmetric tensor $\varepsilon^{ss'} = \varepsilon_{\dot{s}\dot{s}'} = i\sigma^2$ (with $\varepsilon^{12} = +1$) raises and lowers indices, exactly as the metric raises and lowers spacetime indices. We will not need index raising/lowering explicitly until we write Lorentz-invariant products of spinors, but it is good to know the structure is there.

Component form of the general transformation law. A general (j_L, j_R) field has components $\Phi^s_{\dot{s}}(X)$, where $s \in \{1, \dots, 2j_L + 1\}$ is the undotted (left) index and $\dot{s} \in \{1, \dots, 2j_R + 1\}$ is the dotted (right) index. The matrix equation $\Phi(X) \rightarrow D^{(j_L)}(L) \Phi(X) D^{(j_R)}(L^\dagger)$ becomes, in components:

$$\Phi^s_{\dot{s}}(X) \mapsto D^{(j_L)}(L)^s_{s'} \Phi^{s'}_{\dot{s}'}(\Lambda^{-1}X) D^{(j_R)}(\bar{L})^{\dot{s}'}_{\dot{s}}. \quad (26.43)$$

Reading this equation carefully:

- s' and $\dot{s}'$ are *contracted* (summed) indices, running over the internal spinor dimensions. They are the “dummy” indices of Einstein summation.
- s and $\dot{s}$ are *free* indices, labelling which component of the transformed field we are computing. They appear once on the left and once on the right.
- The undotted free index s appears as an upper index on both Φ and $D^{(j_L)}(L)$, confirming it transforms in the left space.
- The dotted free index $\dot{s}$ appears as a lower index on both Φ and $D^{(j_R)}(\bar{L})$, confirming it transforms in the right space.
- The argument shifts from X to $\Lambda^{-1}X$ because the field is evaluated at the *original* spacetime point expressed in the new frame.

Special cases. The general law (26.43) reduces to the following for each particle type:

$$\text{Scalar: } \phi(X), \quad \text{no indices,} \quad \phi(X) \rightarrow \phi(\Lambda^{-1}X). \quad (26.44)$$

$$\begin{aligned} \text{Left Weyl spinor: } \psi^s(X), \quad s \in \{1, 2\}, \quad \text{no dotted index,} \\ \psi^s(X) \mapsto L^s_{s'} \psi^{s'}(\Lambda^{-1}X). \end{aligned} \quad (26.45)$$

$$\begin{aligned} \text{Right Weyl spinor: } \bar{\chi}^{\dot{s}}(X), \quad \dot{s} \in \{1, 2\}, \quad \text{no undotted index,} \\ \bar{\chi}^{\dot{s}}(X) \mapsto \bar{L}^{\dot{s}}_{\dot{s}'} \bar{\chi}^{\dot{s}'}(\Lambda^{-1}X). \end{aligned} \quad (26.46)$$

$$\begin{aligned} \text{4-vector: } A^s_{\dot{s}}(X) &= A_\mu(X) (\sigma^\mu)^s_{\dot{s}}, \\ A^s_{\dot{s}}(X) &\mapsto L^s_{s'} A^{s'}_{\dot{s}'}(\Lambda^{-1}X) \bar{L}^{\dot{s}'}_{\dot{s}}. \end{aligned} \quad (26.47)$$

The 4-vector case shows clearly why σ^μ carries one undotted upper index and one dotted lower index: it is the canonical $(\frac{1}{2}, \frac{1}{2})$ object that converts a spacetime index μ into one index from each spinor space.

Lorentz-invariant contractions. The index notation makes it mechanical to write Lorentz scalars (quantities invariant under (26.43)). The rule is the same as in tensor calculus: *contract an upper index with a lower index of the same type*. Concretely:

- Two undotted upper indices cannot be contracted directly; one must be lowered with $\varepsilon_{ss'}$ first.
- An undotted index can never be contracted with a dotted index.
- A Lorentz scalar has no free indices of any kind.

The canonical Lorentz-invariant spinor products are therefore:

$$\psi\chi \equiv \varepsilon_{ss'}\psi^s\chi^{s'} = \psi^1\chi^2 - \psi^2\chi^1, \quad \bar{\psi}\bar{\chi} \equiv \varepsilon_{\dot{s}\dot{s}'}\bar{\psi}^{\dot{s}}\bar{\chi}^{\dot{s}'}, \quad (26.48)$$

and the cross-chirality pairing (which will become the mass term in the Dirac Lagrangian):

$$\bar{\chi}_{\dot{s}}(\bar{\sigma}^\mu)^{\dot{s}s}\psi_s \equiv \bar{\chi}\bar{\sigma}^\mu\psi. \quad (26.49)$$

Every term in the Standard Model Lagrangian is built from such contractions, which is why the dot/no-dot notation pays for itself almost immediately once one starts reading the actual literature.

Remark. [ML analogy: typed tensor contraction] In a typed tensor library (e.g. `einops` or named-tensor frameworks), you are not allowed to contract two axes unless they carry the same type label. Van der Waerden notation enforces exactly the same discipline: undotted and dotted indices are different types, and the type system prevents illegal contractions at the level of notation rather than at runtime. Every Lorentz-invariant expression in spinor notation is a well-typed contraction.

26.10 Indices vs. labels: what transforms and what does not

Having introduced both spinor indices ($s, \dot{s}$) and chirality labels (L, R), it is essential to be precise about which of these objects the Lorentz group actually *acts on*. The confusion between the two is one of the most common sources of error when first reading QFT literature.

Spinor indices are slots; the group acts on them. The undotted index $s \in \{1, 2\}$ and the dotted index $\dot{s} \in \{1, 2\}$ are *numerical slots* inside a field component $\Phi^s_{\dot{s}}(X)$. A Lorentz transformation L genuinely *moves values between these slots*: the component $\Phi^1_{\dot{1}}$ mixes with $\Phi^2_{\dot{1}}$, and so on, according to equation (26.43). This is what it means for a field to “transform under the Lorentz group”: the numerical values stored in the slots get linearly reshuffled by the matrix L (for undotted slots) or $\bar{L}$ (for dotted slots).

Concretely, for a left Weyl spinor with components $\psi^s = (\psi^1, \psi^2)^T$:

$$\begin{pmatrix} \psi^1 \\ \psi^2 \end{pmatrix} \mapsto \begin{pmatrix} L^1_1 & L^1_2 \\ L^2_1 & L^2_2 \end{pmatrix} \begin{pmatrix} \psi^1 \\ \psi^2 \end{pmatrix}. \quad (26.50)$$

The index $s = 1$ or $s = 2$ tells you *which slot* you are reading; the matrix L tells you how those slots mix. Both pieces of information are necessary, and they are conceptually distinct.

The labels L and R are names, not slots. By contrast, the subscripts L and R in ψ_L and ψ_R are *human-assigned names*. They record which representation the field was *declared to belong to* when it was introduced:

- ψ_L was declared to be a $(\frac{1}{2}, 0)$ field. The label L is a reminder of this declaration.
- ψ_R was declared to be a $(0, \frac{1}{2})$ field. The label R is a reminder of this declaration.

No matrix ever acts on the label itself. The label does not run over values $\{1, 2\}$; it does not appear in a summation; it is never contracted. It is purely documentary, like a variable name in a programming language:

$$\text{psi}_L \neq \text{psi}_R \quad \text{are two different fields, not two components of one field.} \quad (26.51)$$

A Lorentz transformation does not turn ψ_L into ψ_R or vice versa. The representation type is fixed at the time the field is defined; no physical process described by a Lorentz transformation alone can change it.

The label determines which index type appears. The connection between the two is:

Label	Representation	Index type carried
L	$(\frac{1}{2}, 0)$	undotted index s , transforms with L
R	$(0, \frac{1}{2})$	dotted index $\dot{s}$, transforms with $\bar{L}$

The label L on the field ψ_L tells you that its components carry undotted indices, written ψ^s . The label R on ψ_R tells you its components carry dotted indices, written $\bar{\psi}^{\dot{s}}$. Once you see the index, the label has done its job and could in principle be dropped—the index type alone encodes the representation. In practice, both the label and the index are kept for readability.

Dirac spinor: the label is a block address. In the Dirac spinor

$$\Psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} = \begin{pmatrix} \psi^1 \\ \psi^2 \\ \bar{\psi}^{\dot{1}} \\ \bar{\psi}^{\dot{2}} \end{pmatrix}, \quad (26.52)$$

the labels L and R now serve as *block addresses*: they tell you that the upper two slots carry undotted indices and the lower two carry dotted indices. The 4×4 Lorentz matrix $S(L) = \text{diag}(L, \bar{L})$ respects this block structure. The label is still not an index; it is still not summed; it is still not acted on. It is a structured comment telling you the type of each block.

Remark. [ML analogy: **dtype** vs. **value**] In a tensor library, every tensor has a **dtype** (e.g. `float32`, `complex64`) that is fixed when the tensor is created and does not change when you apply a linear map to the tensor’s values. The spinor label L or R is the **dtype**: it records what kind of object this is. The spinor index s or $\dot{s}$ is the axis index: it selects a specific entry along a specific axis. A matrix multiplication (Lorentz transformation) changes the *values* stored at each index, but it does not change the *dtype* of the tensor. You cannot accidentally turn a `float32` tensor into a `complex64` tensor by multiplying by a matrix, just as you cannot turn a left-chiral field into a right-chiral field by applying a Lorentz transformation.

26.11 From equivariant fields to an invariant Lagrangian

We have now constructed the equivariant fields of the Standard Model. Let us take stock of what we have built, before assembling the pieces into a Lagrangian.

At each spacetime point x , a field assigns a local variable: a number, a vector, or a spinor. The type of local variable determines how it transforms when we change reference frame. A scalar field $\phi(x)$ assigns a single number — it does not care about orientation or velocity. A vector field $A^\mu(x)$ assigns a four-component arrow that rotates and boosts with the frame. A left-handed Weyl spinor field $(\psi_L)_s(x)$ assigns a two-component object transforming under the $(\frac{1}{2}, 0)$ representation — it is sensitive to rotations in a subtler way than a vector, acquiring a sign under a full 2π rotation. A right-handed Weyl spinor field $(\psi_R)^{\dot{s}}(x)$ transforms under the $(0, \frac{1}{2})$ representation: the conjugate of the left-handed one. A Dirac spinor $\Psi_A(x)$ combines both into a four-component object:

$$\Psi_A = \begin{pmatrix} (\psi_L)_s \\ (\psi_R)^{\dot{s}} \end{pmatrix}, \quad (26.53)$$

transforming under $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$. In every case, the transformation rule is linear and consistent: applying two successive Lorentz transformations is the same as applying their composition. This is the equivariance condition (26.4), and it is the only constraint we imposed on the fields in this section.

The end game is an invariant Lagrangian. The question is: given a collection of equivariant fields, each carrying its own type of index, how do we build a Lorentz scalar $\mathcal{L}$ from them? The answer is purely algebraic: contract all free Lorentz and spinor indices using the invariant tensors of the representations involved, so that every index is absorbed and the result transforms trivially — that is, does not transform at all.

The general principle. If Φ_1 transforms in representation ρ_1 and Φ_2 transforms in representation ρ_2 , their product $\Phi_1 \otimes \Phi_2$ transforms in the tensor product representation $\rho_1 \otimes \rho_2$. By the Clebsch-Gordan decomposition, this tensor product decomposes into a direct sum of irreducible representations. A Lorentz scalar is extracted by projecting onto the trivial (singlet) component: the component that transforms as $\rho(\Lambda) = 1$ for all Λ . The invariant tensor that performs this projection is the *contraction* specific to each pair of representations.

Contracting vector indices: the Minkowski metric. Two four-vectors V^μ and W^ν , each transforming as $V^\mu \rightarrow \Lambda^\mu{}_\nu V^\nu$, combine to a scalar via contraction with the Minkowski metric:

$$V^\mu W_\mu = \eta_{\mu\nu} V^\mu W^\nu \mapsto \eta_{\mu\nu} (\Lambda^\mu{}_\rho V^\rho) (\Lambda^\nu{}_\sigma W^\sigma) = \eta_{\rho\sigma} V^\rho W^\sigma, \quad (26.54)$$

where the last step uses $\Lambda^T \eta \Lambda = \eta$. The metric $\eta_{\mu\nu}$ is the invariant tensor of the vector representation. Examples:

$$\partial_\mu \phi \partial^\mu \phi, \quad A_\mu A^\mu, \quad F_{\mu\nu} F^{\mu\nu}, \quad \bar{\Psi} \gamma^\mu \partial_\mu \Psi. \quad (26.55)$$

Contracting left-handed Weyl spinor indices: the spinor metric. Two left-handed Weyl spinors $(\psi_L)_s$ and $(\psi'_L)_{s'}$, each transforming as $(\psi_L)_s \rightarrow M_s{}^{s'} (\psi_L)_{s'}$ with $M \in SL(2, \mathbb{C})$, combine to a scalar via contraction with the antisymmetric spinor metric $\epsilon^{ss'}$:

$$\epsilon^{ss'} (\psi_L)_s (\psi'_L)_{s'} = (\psi_L)^s (\psi'_L)_s \mapsto \epsilon^{ss'} (M_s{}^t (\psi_L)_t) (M_{s'}{}^{t'} (\psi'_L)_{t'}) = \epsilon^{tt'} (\psi_L)_t (\psi'_L)_{t'}, \quad (26.56)$$

where the last step uses $\det M = 1$: $\epsilon^{ss'} M_s{}^t M_{s'}{}^{t'} = \epsilon^{tt'}$. The spinor metric $\epsilon^{ss'}$ is the invariant tensor of the $(\frac{1}{2}, 0)$ representation. Similarly, two right-handed Weyl spinors $(\psi_R)^{\dot{s}}$ and $(\psi'_R)^{\dot{s}'}$ contract with $\epsilon_{\dot{s}\dot{s}'}$:

$$\epsilon_{\dot{s}\dot{s}'} (\psi_R)^{\dot{s}} (\psi'_R)^{\dot{s}'} \quad (26.57)$$

Contracting left-handed with right-handed: the sigma matrices. A left-handed Weyl spinor $(\psi_L)_s$ and a right-handed Weyl spinor $(\psi_R)^{\dot{s}}$ do not directly contract to a scalar: their representations are conjugate but not identical. However, together with a spacetime index, the sigma matrices $\sigma^\mu_{s\dot{s}}$ provide the intertwiner:

$$(\psi_L)^s \sigma^\mu_{s\dot{s}} (\psi_R)^{\dot{s}} \quad (26.58)$$

transforms as a four-vector V^μ , which can then be contracted with another vector index via $\eta_{\mu\nu}$. The sigma matrices are the Clebsch-Gordan coefficients that decompose $(\frac{1}{2}, 0) \otimes (0, \frac{1}{2}) = (\frac{1}{2}, \frac{1}{2})$: they convert a pair of Weyl spinors of opposite handedness into a vector.

The Dirac kinetic term: all indices contracted. The Dirac kinetic term:

$$\bar{\Psi} i \gamma^\mu \partial_\mu \Psi = \bar{\Psi}^A i (\gamma^\mu)_{AB} \partial_\mu \Psi^B \quad (26.59)$$

contracts the spinor index A with B via the gamma matrix $(\gamma^\mu)_{AB}$, and then the vector index μ with ∂_μ . In terms of Weyl components (26.53), this decomposes into:

$$\bar{\Psi} i \gamma^\mu \partial_\mu \Psi = i (\psi_R)^{\dot{s}} \bar{\sigma}^{\mu \dot{s}s} \partial_\mu (\psi_L)_s + i (\psi_L)_s^\dagger \sigma^{\mu s\dot{s}} \partial_\mu (\psi_R)_{\dot{s}}, \quad (26.60)$$

showing that the Dirac kinetic term couples left-handed and right-handed Weyl spinors through the sigma matrices: the two handedness sectors communicate via σ^μ .

The mass term. The Dirac mass term:

$$m\bar{\Psi}\Psi = m[(\psi_R)^{\dot{s}}(\psi_L)_s \epsilon_{s\dot{s}} + (\psi_L)^\dagger_s (\psi_R)^\dagger_{\dot{s}} \epsilon^{s\dot{s}}] \quad (26.61)$$

explicitly mixes left-handed and right-handed Weyl spinors: a Dirac mass is only possible when both ψ_L and ψ_R are present. A *Majorana* mass term contracts two left-handed Weyl spinors via $\epsilon^{ss'}$ directly:

$$\frac{m}{2} \epsilon^{ss'} (\psi_L)_s (\psi_L)_{s'} + \text{h.c.}, \quad (26.62)$$

using (26.56): this is possible for a neutral fermion (such as a neutrino in some beyond-Standard-Model extensions) where ψ_L and ψ_R are related by charge conjugation.

The Yukawa coupling. The Yukawa coupling between a Dirac fermion and the Higgs scalar ϕ :

$$y \bar{\Psi} \phi \Psi = y \phi [(\psi_R)^{\dot{s}}(\psi_L)_s \epsilon_{s\dot{s}} + \text{h.c.}] \quad (26.63)$$

is a scalar because ϕ carries no Lorentz or spinor indices, and the remaining ψ_L - ψ_R contraction is a scalar by (29.8). In the Standard Model, ϕ is the Higgs doublet; the Yukawa coupling gives masses to all quarks and charged leptons after the Higgs condenses to $\langle \phi \rangle = v$.

The complete Standard Model Lagrangian as a scalar. Every term in the Standard Model Lagrangian is built by this procedure: assign each field its equivariant representation, take products of fields and their derivatives, and contract all Lorentz and spinor indices with the available invariant tensors ($\eta_{\mu\nu}$, $\epsilon^{ss'}$, $\epsilon_{\dot{s}\dot{s}'}$, $\sigma^\mu_{s\dot{s}}$, $(\gamma^\mu)_{AB}$). The requirement that all indices be contracted is highly restrictive: it limits the allowed interaction terms and forces the specific structure of the Standard Model vertices.

Once the scalar Lagrangian $\mathcal{L}$ is in hand, the round trip (26.6) proceeds:

- **Legendre transform:** $\mathcal{L} \rightarrow H$, singling out time and obtaining the Hamiltonian as a polynomial of fields and conjugate momenta.
- **Mode expansion and quantization:** $H \rightarrow H[a^\dagger, a]$, promoting Fourier coefficients to Fock operators. The Fock basis vectors $|n; \mathbf{k}_j, s_j, \dots\rangle$ inherit their spinor index s_j from the equivariant representation: left-handed particles carry undotted indices s ; right-handed particles carry dotted indices $\dot{s}$.
- **Time slicing:** $e^{-iHT} \rightarrow \int \mathcal{D}\Phi e^{iS}$, inserting Fock resolutions of the identity between each time step. For ψ_L and ψ_R , the conjugate variables at each time slice are Grassmann numbers; the path integral is over Grassmann fields.
- **Dyson series:** expanding $e^{-iH_{\text{int}}T}$ gives Feynman diagrams. Each vertex is one term from $\mathcal{L}_{\text{int}}$ with all indices contracted. The propagator for a Dirac fermion is a 4×4 matrix in spinor space; for a Weyl fermion it is 2×2 ; for a photon it is 4×4 in vector space.

Field	Representation	Index	Invariant tensor
Scalar ϕ	$(0, 0)$	None	Trivial: ϕ^2 , $ \phi ^2$
Left Weyl $(\psi_L)_s$	$(\frac{1}{2}, 0)$	$s \in \{1, 2\}$	$\epsilon^{ss'}$
Right Weyl $(\psi_R)^{\dot{s}}$	$(0, \frac{1}{2})$	$\dot{s} \in \{1, 2\}$	$\epsilon_{\dot{s}\dot{s}'}$
Dirac Ψ_A	$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	$A \in \{1, 2, 3, 4\}$	$(\gamma^\mu)_{AB}$, δ_{AB}
Vector A^μ	$(\frac{1}{2}, \frac{1}{2})$	$\mu \in \{0, 1, 2, 3\}$	$\eta_{\mu\nu}$

The equivariant representation is the input; the invariant Lagrangian is the output. The representation theory of the Lorentz group gives us the complete list of equivariant fields; the invariant tensor algebra tells us how to combine them into scalars. Together, they determine what a Lorentz-invariant quantum field theory can look like — and the Standard Model is, in this precise sense, the minimal such theory consistent with gauge invariance, renormalizability, and the observed particle content.

26.12 Wigner’s classification: what a particle IS

The previous subsections developed the representations of the Lorentz group from the algebra. We worked out the representations and then identified particles with them. Wigner (1939) did something deeper: he showed that the identification is not a convenience but a *definition*. A particle IS an irreducible unitary representation of the Poincaré group. Not “is described by,” not “transforms as”—IS.

This subsection is the intellectual account of how Wigner arrived at this identification and what it means.

The problem Wigner faced

By the late 1930s, quantum mechanics and special relativity were both established, but their marriage was incomplete. The Dirac equation (1928) described the electron beautifully—spin-1/2, antiparticles, the gyromagnetic ratio—but it was one equation for one particle. The question remained: what are the *possible* particles that can exist in a relativistic quantum theory?

The question is not about dynamics (what forces act, what interactions occur). It is about *kinematics*: before writing any Hamiltonian, before specifying any interaction, what are the possible one-particle states that are consistent with the symmetries of spacetime?

Wigner’s insight: the symmetries of spacetime are encoded in the Poincaré group (the Lorentz group $SO(3,1)$ plus spacetime translations $\mathbb{R}^{3,1}$). A one-particle state must transform under the Poincaré group in a consistent way: a Lorentz boost or a rotation or a translation must map one valid one-particle state to another. The set of all one-particle states, together with the action of the Poincaré group on them, is a **representation** of the Poincaré group.

For the representation to describe a single particle (not a multi-particle system or a superposition of different species), it must be **irreducible**: it cannot be decomposed into two independent subspaces that do not mix under the group action. And for the quantum mechanical probabilities to be preserved under symmetry transformations, the representation must be **unitary**: the group elements act as unitary operators on the Hilbert space.

Therefore: the possible one-particle species in a relativistic quantum theory are the **irreducible unitary representations of the Poincaré group**. Classifying these representations classifies the possible particles.

The Casimirs: mass and spin

The Poincaré group has two Casimir operators—operators that commute with every generator of the group, and therefore have a fixed value on each irreducible representation:

The first Casimir: $P^2 = P_\mu P^\mu = -m^2 c^2$.

The squared four-momentum. Its eigenvalue is $-m^2 c^2$, where m is the **mass** of the particle. The mass is Lorentz-invariant (every observer agrees on its value), as it must be for a Casimir.

The mass divides particles into three classes:

- $m > 0$ (**massive**): the four-momentum is timelike ($P^2 < 0$ in the $(-, +, +, +)$ convention). The particle has a rest frame. Examples: electron, proton, Higgs boson.
- $m = 0$ (**massless**): the four-momentum is lightlike ($P^2 = 0$). The particle has no rest frame; it moves at the speed of light. Examples: photon, graviton, gluon (before confinement).
- $m^2 < 0$ (**tachyonic**): the four-momentum is spacelike. No physical particles have been observed in this class; tachyonic fields signal instabilities (the Higgs field before symmetry breaking is tachyonic in this sense).

The second Casimir: $W^2 = W_\mu W^\mu$.

The squared Pauli–Lubanski pseudovector, where:

$$W^\mu = \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} J_{\nu\rho} P_\sigma, \quad (26.64)$$

with $J_{\nu\rho}$ the Lorentz generators and P_σ the translation generators. For a massive particle in its rest frame ($P^\mu = (mc, \mathbf{0})$): $W^0 = 0$ and $W^i = mc J^i$, where J^i are the spatial rotation generators. Therefore $W^2 = -m^2 c^2 \mathbf{J}^2 = -m^2 c^2 s(s+1) \hbar^2$, where s is the **spin** of the particle.

The spin is quantized: $s = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots$. Each value labels a different irreducible representation.

Together: an irreducible representation of the Poincaré group is labelled by two numbers: the mass m and the spin s (for massive particles) or the helicity λ (for massless particles). This is the classification of particles.

Wigner’s method: the little group

Wigner’s construction of the representations uses a beautiful technique: the **method of induced representations**, based on the **little group** (also called the **stabiliser** or the **small group**).

The idea: pick a “standard” four-momentum k^μ (a reference momentum for each class of particle). The **little group** of k^μ is the subgroup of the Lorentz group that leaves k^μ unchanged:

$$\text{Little}(k) = \{W \in SO(3,1) : W^\mu{}_\nu k^\nu = k^\mu\}. \quad (26.65)$$

The little group is the group of “internal” transformations of the particle at rest (or at the reference momentum): the transformations that change the particle’s internal state (spin, polarisation) without changing its momentum.

The full representation is then “induced” from the little group: any four-momentum p^μ with $p^2 = k^2$ can be reached from k^μ by a Lorentz boost $L(p)$ (the “standard boost”: $p^\mu = L(p)^\mu{}_\nu k^\nu$), and the particle’s state at momentum p is obtained by boosting the state at k .

The classification of particles therefore reduces to the classification of representations of the little group.

Massive particles: the little group is $SO(3)$

For a massive particle: the standard four-momentum is the rest-frame momentum $k^\mu = (mc, \mathbf{0})$. The Lorentz transformations that leave this unchanged are precisely the *spatial rotations*: $W^0_0 = 1$, $W^0_i = W^i_0 = 0$, $W^i_j = R^i_j$ (a 3×3 rotation matrix). The little group is:

$$\text{Little}(mc, \mathbf{0}) = SO(3). \quad (26.66)$$

The irreducible representations of $SO(3)$ are labelled by $s = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots$, with dimension $2s + 1$. Each representation is the spin: $s = 0$ (scalar, one state), $s = \frac{1}{2}$ (spinor, two states: spin

up and down), $s = 1$ (vector, three states: three polarisations), $s = 2$ (tensor, five states), and so on.

A massive particle of spin s has $2s + 1$ internal states. These are the spin projections $m_s = -s, -s + 1, \dots, s - 1, s$ —the states that are mixed by rotations but not by boosts (the spin is defined in the rest frame).

The physical content. When we say “the electron has spin $1/2$,” we mean: the electron’s one-particle states, at any momentum $\mathbf{p}$, form a two-dimensional representation of the rotation group $SO(3)$ (the little group of the electron’s rest-frame momentum). The two states ($\uparrow, \downarrow$) are the two spin projections. Rotations mix them; boosts do not (in the rest frame).

The Dirac equation, the Pauli equation, the Weyl equation—all the wave equations that describe particles of definite spin—are simply the *field equations* that enforce the irreducibility condition: they project out the unwanted components and ensure that the field transforms as a single irreducible representation of the little group.

Massless particles: the little group is $ISO(2) \supset SO(2)$

For a massless particle: the standard four-momentum is $k^\mu = (\omega/c, 0, 0, \omega/c)$ (a photon moving in the z -direction, say). The little group is larger than one might expect:

$$\text{Little}(\omega, 0, 0, \omega) = ISO(2) = SO(2) \ltimes \mathbb{R}^2, \quad (26.67)$$

the Euclidean group of the two-dimensional plane perpendicular to the direction of motion. It consists of:

- $SO(2)$: rotations in the transverse (x, y) -plane. These are the physical transformations. The angle of rotation is θ , and the eigenvalue of the generator is the **helicity** λ : the component of angular momentum along the direction of motion.
- $\mathbb{R}^2$: two “translation-like” generators that correspond to certain combinations of boosts and rotations. In all known physical representations: these generators act trivially (their eigenvalues are zero). If they acted nontrivially, the representation would have a continuous internal degree of freedom—a “continuous spin” particle, which has never been observed.

For physical massless particles (with the $\mathbb{R}^2$ generators acting trivially): the little group reduces effectively to $SO(2)$, whose irreducible representations are one-dimensional, labelled by the helicity λ (an integer or half-integer):

$$e^{i\theta J_z} |p, \lambda\rangle = e^{i\lambda\theta} |p, \lambda\rangle. \quad (26.68)$$

The physical content.

A massless particle of helicity λ has *one* internal state at each momentum (not $2|\lambda| + 1$, as for a massive particle of spin $|\lambda|$). The photon ($\lambda = \pm 1$) has two helicity states (right-circular and left-circular polarisation), not three (the longitudinal polarisation is absent because $m = 0$). The graviton ($\lambda = \pm 2$) has two helicity states, not five.

The “missing” polarisations (the longitudinal and scalar components) are absent because the little group is $SO(2)$, not $SO(3)$: there is no rest frame for a massless particle, and the little group of a lightlike momentum is smaller than the little group of a timelike momentum. This is why massless gauge bosons have only two polarisations, and why the photon’s longitudinal mode must be “eaten” by the Higgs mechanism to give the W and Z bosons their mass and their third polarisation.

The difference between massive and massless particles is not merely quantitative (one has $m > 0$, the other $m = 0$); it is *structural*: they have different little groups ($SO(3)$ vs. $SO(2)$), different numbers of internal states ($2s + 1$ vs. 2 for $|\lambda| \geq 1$), and different representation theory.

The classification, summarised

Class	P^2	Standard k^μ	Little group	Label	Examples
Massive	$-m^2 c^2 < 0$	$(mc, \mathbf{0})$	$SO(3)$	Spin s ($2s + 1$ states)	e ($s = \frac{1}{2}$), W ($s = 1$), H ($s = 0$)
Massless	0	$(\omega, 0, 0, \omega)$	$ISO(2)$ $SO(2)$	$\rightarrow$ Helicity λ (2 states)	γ ($\lambda = \pm 1$), graviton ($\lambda = \pm 2$)
Tachyonic	> 0	$(0, 0, 0, \mu c)$	$SO(2, 1)$	Continuous	(Not observed)

Wigner’s classification is exhaustive: every irreducible unitary representation of the Poincaré group falls into one of these classes. Every particle in nature—every excitation of any quantum field theory—belongs to exactly one class, labelled by its mass and spin (or helicity).

What Wigner’s classification means

Wigner’s result is deeper than a catalogue. It has several profound implications.

1. A particle is not a “thing.” It is a representation.

Before Wigner: a particle was imagined as a tiny object with intrinsic properties (mass, charge, spin) that moves through spacetime. The properties were attributes of the object, like the colour or weight of a marble.

After Wigner: a particle is an *irreducible representation* of the Poincaré group. Its mass and spin are not attributes of an object; they are *labels of a representation*—the Casimir eigenvalues that classify the representation. The “particle” is the representation; the representation is the particle.

This is a shift in ontology: from “a thing with properties” to “a pattern of transformation.” The electron is not a tiny ball that happens to have $s = 1/2$; it is the unique spin-1/2 representation of the Poincaré group, realised in nature. Its properties (mass, spin, number of internal states, behaviour under boosts and rotations) are not contingent features of an object but necessary consequences of the representation theory.

2. The classification is kinematic, not dynamic.

Wigner’s classification uses only the symmetries of spacetime (the Poincaré group). It does not involve any Hamiltonian, any Lagrangian, any interaction, or any dynamical law. The possible particles are determined *before* the dynamics is specified.

This means: the spectrum of possible particles is the same for every relativistic quantum theory, regardless of the interactions. QED, QCD, the Standard Model, any beyond-the-Standard-Model theory—all must build their particle content from the same menu of Poincaré representations. The dynamics determines *which* representations are realised (which particles exist) and how they interact, but the *menu* of possibilities is fixed by the symmetry.

3. Spin is not angular momentum of a spinning object.

The classical picture: the electron “spins” on its axis, and the spin angular momentum is the angular momentum of this rotation. This picture is wrong (the electron has no axis, no surface, no internal structure to rotate), but the language persists.

Wigner’s classification reveals what spin actually is: it is the **label of the little group representation**. For a massive particle: s labels the $SO(3)$ representation (how the particle’s internal states transform under rotations in the rest frame). For a massless particle: λ labels the $SO(2)$ representation (how the particle’s state transforms under rotations around the direction of motion).

The spin is not a property of a rotating object. It is a property of a representation: the number of internal states and the rule for mixing them under rotations. A spin-1/2 particle has two states that rotate into each other under $SU(2)$; a spin-1 particle has three states that rotate into each other under $SO(3)$. The “rotation” is a mathematical operation on the state space, not a physical rotation of an extended body.

4. Massless particles are qualitatively different from massive ones.

The little group for massive particles is $SO(3)$ (compact, three-dimensional, with finite-dimensional representations labelled by s). The little group for massless particles is $ISO(2)$ (non-compact, with a different structure entirely).

This structural difference is responsible for: the absence of longitudinal polarisations for massless gauge bosons, the need for gauge invariance (to project out the unphysical longitudinal and scalar polarisations from the covariant field), the Higgs mechanism (which changes the little group from $ISO(2)$ to $SO(3)$ by giving the gauge boson a mass), and the impossibility of giving a photon a mass without breaking gauge invariance.

All of these deep physical facts are consequences of representation theory—of the difference between the little groups $SO(3)$ and $ISO(2)$.

Wigner and the game engine

In the language of this book: Wigner’s classification is the classification of the **Fock-layer degrees of freedom**.

The Fock space is built from creation operators $\hat{a}_{\mathbf{k},\lambda}^\dagger$, where $\mathbf{k}$ is the momentum and λ is the little-group label (spin projection for massive particles, helicity for massless particles). Each creation operator creates a one-particle state that transforms as an irreducible representation of the Poincaré group. The Fock space is the direct sum of n -particle sectors, each built by applying n creation operators to the vacuum.

Wigner’s classification determines *what the game engine’s registers can hold*: each register stores a particle characterised by its mass m (the P^2 Casimir) and its spin s or helicity λ (the W^2 Casimir). The Hamiltonian is a polynomial in these operators—the dynamics of the game engine. The Born rule is the readout.

The classification is the *input specification* of the game engine: the format of the data that the engine can process. It does not tell you what the engine does (that is the Hamiltonian); it tells you what the engine’s variables are. And the variables are not arbitrary: they are constrained by the symmetry of the engine’s internal grid ($\eta_{\mu\nu}$ and the Poincaré group).

Wigner’s 1939 paper is, in this light, the specification document for the quantum game engine: a complete list of the possible data types (mass and spin), the allowed operations (Lorentz boosts and rotations), and the constraints on the data (irreducibility and unitarity). Everything that follows—QED, QCD, the Standard Model, the Higgs mechanism, the running of coupling constants—is the dynamics, written in the language of these data types.

27 Adding an Internal Indices for Gauge Transformation

27.1 Internal indices and local gauge invariance

In the previous section, we enriched the field variable at each spacetime point x by adding spinor indices s and $\dot{s}$. These indices transform under matrix representations of the Lorentz group: when we change reference frame by a Lorentz transformation Λ , the spinor indices of the field variable rotate by the corresponding representation matrix. The field is an equivariant map from spacetime to its representation space, and the requirement that the Lagrangian be Lorentz invariant forces us to contract all spinor and spacetime indices with the appropriate invariant tensors.

We now add one more layer of structure.

Adding an internal index. At each spacetime point x , and for each spinor index $(s, \dot{s})$, we attach an additional index $i \in \{1, \dots, N\}$. The field variable at x is now a three-way object: it carries a spacetime location, a spinor index, and an internal index. In the quark sector of QCD, for example, $i \in \{1, 2, 3\}$ labels the three colour charges (red, green, blue); in the electroweak sector, $i \in \{1, 2\}$ labels the isospin doublet (electron and neutrino, or up quark and down quark). The internal index has nothing to do with spacetime geometry: it lives in an abstract internal space attached to each spacetime point, independent of position, spin, or momentum.

An $N \times N$ unitary matrix $U \in SU(N)$ can act on this internal index, rotating the N components of the field into each other. This is an internal symmetry transformation: it does not move the spacetime point x , it does not touch the spinor index, it only rotates the internal components. If the same matrix U is applied at every spacetime point simultaneously — a *global* gauge transformation — this is a straightforward generalisation of the global symmetries familiar from the previous section. The Lagrangian can be made invariant under global $SU(N)$ transformations by contracting the internal indices with the appropriate $SU(N)$ -invariant tensors (the Kronecker delta δ_{ij} and the structure constants f^{abc}), in exact analogy with contracting spinor indices with $\epsilon^{ss'}$.

Where things get interesting: local gauge transformations. Now allow the matrix $U(x)$ to depend on the spacetime point x . At each point, a different $SU(N)$ rotation is applied to the internal index. This is a *local* gauge transformation. It is a much larger symmetry: instead of one matrix U for the entire universe, there is an independent matrix $U(x)$ at each point, giving an infinite-dimensional group of transformations.

A Lagrangian built from spinor fields alone is *not* invariant under local gauge transformations, even if it is invariant under global ones. The reason is the derivative: the kinetic term $\partial_\mu \psi_i$ involves comparing the field at two nearby spacetime points x and $x + dx$, but if $U(x)$ and $U(x + dx)$ are different matrices, the derivative does not transform covariantly — it picks up an extra term proportional to $\partial_\mu U(x)$. This spoils the invariance of the Lagrangian.

The resolution requires introducing a new field — the *gauge field* $A_\mu^a(x)$ — specifically designed to compensate for the variation of $U(x)$ from point to point. The gauge field lives in the Lie algebra of $SU(N)$, carries both a spacetime vector index μ and an internal adjoint index a , and transforms in a specific way under local gauge transformations that exactly cancels the extra term from $\partial_\mu U(x)$. The combination $D_\mu \psi_i = \partial_\mu \psi_i - ig A_\mu^a T_{ij}^a \psi_j$ — the *covariant derivative* — transforms covariantly under local gauge transformations, and replacing ∂_μ by D_μ in the Lagrangian restores the invariance. The gauge field A_μ^a is not added by hand; it is *forced* by the requirement of local gauge invariance. This is the origin of all forces in the Standard Model: the photon, the $W^\pm$ and Z^0 , and the gluons are all gauge fields introduced to maintain local gauge invariance of the matter Lagrangian.

Why local gauge invariance? It is natural to ask: Lorentz invariance is forced on us by special relativity, but why should we demand local gauge invariance? The answer, from the Fock-Born perspective, is that local gauge invariance is not an additional physical requirement — it is a *redundancy reduction*.

The gauge field $A_\mu^a(x)$ has more components than there are physical degrees of freedom. A massless vector field in four dimensions has four components, but only two are physical (the two transverse polarisations); the other two are gauge artefacts. Field configurations related by a local gauge transformation $U(x)$ are physically identical: they describe the same physical state. The space of all field configurations is therefore far larger than the space of physical states, and integrating over all configurations in the path integral overcounts each physical state infinitely many times — once for every element of the local gauge group.

Local gauge invariance is the statement that the Lagrangian, and hence the physics, is constant along the *gauge orbits*: the sets of field configurations related by gauge transformations. The physical variables are not the fields themselves but the gauge orbits — the equivalence classes of fields under gauge transformations. This is a dimension reduction: from the enormous space of all field configurations to the much smaller space of gauge orbits. The Faddeev-Popov procedure and BRST cohomology implement this dimension reduction systematically, ensuring that the Fock-Hamiltonian framework is well-defined: the physical Hilbert space $\ker Q / \text{im } Q$ contains only gauge-invariant states, and the Hamiltonian acting on it generates only physically meaningful transitions.

Without local gauge invariance, the path integral would integrate over an infinite redundancy, the Fock space would contain unphysical states with negative norm (ghost states from the longitudinal and temporal polarisations of the gauge field), and the perturbative expansion would give nonsensical results. Local gauge invariance is the structural principle that makes the theory well-behaved.

The Fock basis vectors with internal indices. The Fock basis vectors now carry three types of labels: the spacetime momentum $\mathbf{k}$, the spinor index s or $\dot{s}$ from the Lorentz representation, and the internal index i from the gauge representation. For a quark in QCD, a complete single-particle basis vector is:

$$|\mathbf{k}, s, i\rangle, \quad \mathbf{k} \in \mathbb{R}^3, \quad s \in \{1, 2\}, \quad i \in \{1, 2, 3\}, \quad (27.1)$$

specifying the momentum, the spin, and the colour of the quark. For a gluon, the internal index $a \in \{1, \dots, 8\}$ labels the eight generators of $SU(3)_c$. The Fock space of QCD is the direct sum of all sectors built from these basis vectors, with the Hamiltonian a polynomial of the corresponding creation and annihilation operators, and the physical Hilbert space obtained by projecting onto the gauge-invariant (colour-singlet) subspace.

The structure is the same as the non-relativistic Fock space, with two additions: the spinor index (from Lorentz invariance) and the internal index (from gauge invariance). Both are forced by consistency requirements; neither is put in by hand. The task of the present section is to make this structure precise: to classify the representations of the gauge group, to derive the covariant derivative, to write the gauge-invariant Lagrangian, and to identify the gauge orbits that the Faddeev-Popov procedure will later integrate over.

27.2 Adding one internal index

So far the field $\Psi_s^s(x)$ carries only Lorentz spinor indices. We now add one further index, completely independent of Lorentz structure, called an *internal* index:

$$\Psi_s^i(x), \quad i \in \{1, 2, \dots, N\}. \quad (27.2)$$

The index i labels a degree of freedom that has nothing to do with spacetime symmetry. At each fixed spacetime point x and for each fixed pair of Lorentz spinor indices $(s, \dot{s})$, the values

$$\begin{pmatrix} \Psi_{\dot{s}}^{s,1}(x) \\ \Psi_{\dot{s}}^{s,2}(x) \\ \vdots \\ \Psi_{\dot{s}}^{s,N}(x) \end{pmatrix} \in \mathbb{C}^N \quad (27.3)$$

form an $N \times 1$ column vector in the linear-algebra sense. We call this the *internal slice* of the field at $(x, s, \dot{s})$.

A group transformation acting on the internal index is a linear map on this column vector. Concretely, let U be an $N \times N$ matrix (for example, an element of $SU(N)$). It acts on the internal index by matrix multiplication:

$$\Psi_{\dot{s}}^{s,i}(x) \mapsto U^i_j \Psi_{\dot{s}}^{s,j}(x), \quad (27.4)$$

where j is a contracted (summed) index and i is the free output index. Written out in full for $N = 2$:

$$\begin{pmatrix} \Psi_{\dot{s}}^{s,1} \\ \Psi_{\dot{s}}^{s,2} \end{pmatrix} \mapsto \begin{pmatrix} U^1_1 & U^1_2 \\ U^2_1 & U^2_2 \end{pmatrix} \begin{pmatrix} \Psi_{\dot{s}}^{s,1} \\ \Psi_{\dot{s}}^{s,2} \end{pmatrix}. \quad (27.5)$$

This is ordinary matrix multiplication on the $\mathbb{C}^N$ column vector (27.3). We call i an *N -let index*, and the field is said to transform in the *fundamental representation* of $SU(N)$ if $U \in SU(N)$.

27.3 The field as a 3-way tensor

Once we restore all indices, the field $\Psi_{\dot{s}}^{s,i}(x)$ is a *3-way tensor* in the PyTorch sense: a multi-dimensional array with one axis for each index type.

Axis	Index	Size	Acted on by
Left spinor	$s \in \{1, 2\}$	2	$L \in SL(2, \mathbb{C})$
Right spinor	$\dot{s} \in \{1, 2\}$	2	$\bar{L} \in SL(2, \mathbb{C})$
Internal	$i \in \{1, \dots, N\}$	N	$U \in SU(N)$

In PyTorch terms, this field has shape `[spinor_left=2, spinor_right=2, internal=N]`, and each cell is a complex number.

The full transformation law under a simultaneous Lorentz transformation L and internal transformation U is:

$$\Psi_{\dot{s}}^{s,i}(x) \mapsto L^s_{s'} U^i_j \Psi_{\dot{s}'}^{s',j}(\Lambda^{-1}x) \bar{L}^{\dot{s}'}_{\dot{s}}. \quad (27.6)$$

There are three independent matrix actions here:

- $L^s_{s'}$ acts on the left spinor axis, summing over s' .
- $\bar{L}^{\dot{s}'}_{\dot{s}}$ acts on the right spinor axis, summing over $\dot{s}'$.
- U^i_j acts on the internal axis, summing over j .

The three actions commute with each other because they act on *different axes* of the tensor. In PyTorch notation this is:

```
einsum('ss', dd', ij, s'd'j -> sdi', L, Lbar, U, Psi).
```

The key insight: fixing any two axes and varying the third gives a vector that is acted on by the corresponding matrix. In particular, fixing $(s, \dot{s})$ and varying i gives the internal slice (27.3), on

which U acts, and the *same* matrix U acts on every such slice regardless of the values of $(s, \dot{s})$. The internal transformation is blind to Lorentz structure, and the Lorentz transformation is blind to the internal index.

27.4 Singlets, N -lets, and suppressing the internal index

N -lets. The internal index $i \in \{1, \dots, N\}$ is called an N -let: a doublet for $N = 2$ (acted on by $SU(2)$), a triplet for $N = 3$ (acted on by $SU(3)$), and so on. The terminology reflects the number of components that mix under the internal symmetry.

Singlets and $U(1)$. The special case $N = 1$ deserves separate attention. If the internal index takes only the single value $i = 1$, then the “matrix” U is a 1×1 matrix—just a complex number of modulus one:

$$U = e^{i\alpha} \in U(1), \quad \alpha \in \mathbb{R}. \quad (27.7)$$

The transformation law (27.4) reduces to

$$\Psi_{\dot{s}}^{s,1}(x) \mapsto e^{i\alpha} \Psi_{\dot{s}}^{s,1}(x), \quad (27.8)$$

a simple overall phase rotation. Since i can only take the value 1, it carries no information and is conventionally *suppressed*: we write $\Psi_{\dot{s}}^s(x)$ instead of $\Psi_{\dot{s}}^{s,1}(x)$, with the understanding that the field transforms with a phase $e^{i\alpha}$ under $U(1)$. The $U(1)$ charge (the exponent weight, sometimes called hypercharge Y) is then carried as metadata on the field—a label, not an index—recording how many powers of $e^{i\alpha}$ the field acquires. For example, a field with hypercharge $Y = +\frac{1}{6}$ transforms as $\Psi \mapsto e^{i\alpha/6}\Psi$.

Suppressing spinor indices when they collapse to size 1. An analogous suppression applies to the spinor indices. Recall

- A *left Weyl spinor* ψ_L has only an undotted index $s \in \{1, 2\}$ and no dotted index. The dotted axis is absent (size 0, not size 1), so $\dot{s}$ is suppressed entirely.
- A *right Weyl spinor* ψ_R has only a dotted index $\dot{s} \in \{1, 2\}$ and no undotted index. Similarly s is absent.
- A *scalar* field has neither spinor index; both are absent.

This is different from the $U(1)$ case: for a singlet, the index is present but trivial (size 1); for a Weyl spinor, the other-chirality index is completely absent. Suppression in both cases means the index does not appear in the notation, but for different reasons.

27.5 Suppressing spinor indices without collapsing them

In most QFT writing, the spinor indices $(s, \dot{s})$ are suppressed even when the field is *not* a singlet in spinor space—the field still has a full 2×2 spinor structure, but the indices are simply not written. This is the source of much notational confusion for newcomers, so it is worth being precise.

When we write $\Psi^i(x)$, showing only the internal index i and suppressing $(s, \dot{s})$, we do *not* mean that the spinor axes have been set to size 1. We mean:

The spinor structure is still present in full, but it is being handled implicitly by the gamma matrices and the Dirac conjugation $\bar{\Psi} = \Psi^\dagger \gamma^0$. Any Lorentz-invariant expression built from Ψ automatically contracts the spinor indices correctly via the gamma algebra.

The full transformation law under both Lorentz and internal symmetry, with all indices written out, is equation (27.6). Suppressing spinor indices, this is compressed to:

$$\Psi^i(x) \longmapsto U^i_j \Psi^j(\Lambda^{-1}x), \quad (27.9)$$

where it is understood that Ψ^j still carries spinor indices that transform simultaneously with L and $\bar{L}$.

Index placement for the internal index. The convention for the internal index i follows the same upper/lower logic as the spinor indices:

- An *upper* internal index Ψ^i transforms with U^i_j , i.e., the matrix U acts to produce the output component i .
- A *lower* internal index Ψ_i transforms with $(U^{-1})^j_i = (U^\dagger)^j_i$ (for unitary U), i.e., the conjugate matrix acts.

In the fundamental representation of $SU(N)$, quarks carry an upper colour index q^α and antiquarks carry a lower colour index $\bar{q}_\alpha$, so that the contraction $\bar{q}_\alpha q^\alpha$ is gauge invariant. This upper/lower placement for the internal index is the same convention as for spinor indices, and the rule is the same: contract an upper index with a lower index of the same type to form an invariant.

Note that the internal index i is *not* dotted or undotted—the dot/no-dot notation is reserved exclusively for the two $SL(2, \mathbb{C})$ spinor families. The internal index is distinguished from the spinor indices by its label (i, j, α, β , etc.) and by context.

27.6 Tensors in ML vs. tensors in physics

We have been using the word “tensor” freely, but it means different things in machine learning and in physics. The collision of vocabulary is real and causes genuine confusion. Here we draw the distinction carefully.

Tensor in the ML sense. In PyTorch, JAX, or any modern ML framework, a *tensor* is simply a *multi-dimensional array of numbers*, with no transformation law attached. A tensor of shape $[2, 3, N]$ is a rectangular box of $2 \times 3 \times N$ complex numbers, indexed by three integer indices. Nothing is assumed about how these numbers change when you change coordinates. The word “tensor” here is a generalisation of “matrix” (2-way array) and “vector” (1-way array, i.e., a 1D array of N numbers).

When we write $\Psi^s_i(x)$ and call it a “3-way tensor”, we mean it in this ML sense: it is a multi-dimensional array with three index slots.

Tensor in the physics sense. In physics, a *tensor* is a multi-dimensional array *together with a specific transformation law* under a symmetry group, usually the Lorentz group. The transformation law is part of the definition: without it, the object is just an array, not a tensor.

The canonical example is the *4-vector* $x^\mu = (x^0, x^1, x^2, x^3)$. In the ML sense, this is just a 1D array (column vector) of four real numbers. In the physics sense, it is a *contravariant Lorentz vector*: under a Lorentz transformation Λ , its components transform as

$$x^\mu \longmapsto \Lambda^\mu_\nu x^\nu. \quad (27.10)$$

The upper index μ signals this transformation law. A *covariant* vector (or covector) $x_\mu = \eta_{\mu\nu} x^\nu$ has a lower index and transforms with the inverse transpose:

$$x_\mu \longmapsto (\Lambda^{-1})^\nu_\mu x_\nu. \quad (27.11)$$

A general rank- (p, q) physics tensor $T^{\mu_1 \dots \mu_p}_{\nu_1 \dots \nu_q}$ has p upper (contravariant) and q lower (covariant) Lorentz indices, each transforming with Λ or Λ^{-T} respectively. Contracting an upper index with a lower index (summing over μ) gives an object with fewer free indices that is still a Lorentz tensor—this is the Einstein summation convention.

“Vector” in ML vs. physics. The same distinction applies to the word “vector”:

	ML sense	Physics sense
Vector	1D array of N numbers shape = [N]	4-component object transforming as x^μ upper index μ , transforms with Λ
Dimension	N (any integer)	always 4 (spacetime dimensions)
Column vector	$N \times 1$ array	not standard terminology

When we said earlier that a left Weyl spinor ψ_L is a “ 2×1 column vector”, we meant it in the ML sense: a 1D array of 2 complex numbers. This is emphatically *not* a physics vector (4-vector): a Weyl spinor transforms under $L \in SL(2, \mathbb{C})$, not under the 4×4 Lorentz matrix Λ .

Physics tensors are ML tensors with a constraint. Every physics tensor is an ML tensor (it is an array of numbers), but it comes equipped with a transformation law that the array must satisfy. Conversely, an ML tensor is not automatically a physics tensor—it becomes one only when you specify how its components change under symmetry transformations.

In this book we use the word “tensor” in the ML sense (a multi-dimensional array) unless we explicitly write out the Lorentz index structure (upper/lower $\mu, \nu, \dots$ indices with the Λ transformation law). Spinor indices $s, \dot{s}$ and internal indices i, j always refer to the representation-theoretic transformation laws we have defined above, and are never implicitly “physics tensors” in the μ, ν sense.

Remark. [Summary of index conventions]

Index	Label	Range	Transforms with	Sense
μ, ν	Greek, no dot	$\{0, 1, 2, 3\}$	$\Lambda^\mu{}_\nu$ (Lorentz)	Physics tensor
s	Latin, no dot	$\{1, 2\}$	$L^s{}_{s'}$ (left $SL(2, \mathbb{C})$)	Spinor
$\dot{s}$	Latin, dotted	$\{1, 2\}$	$\bar{L}^{\dot{s}}{}_{\dot{s}'}$ (right $SL(2, \mathbb{C})$)	Spinor
i, j, α, β	Latin	$\{1, \dots, N\}$	$U^i{}_j$ (gauge group)	Internal

In all cases: contracting an upper index with a lower index of the same type gives an invariant quantity. Never contract across types: μ with s , or s with $\dot{s}$, or s with i , are all illegal contractions.

27.7 A note on terminology: “quantum number”

The physics literature uses the term *quantum number* to mean what we have been calling an *index* on a basis vector. The two usages are the same thing.

When we write a basis vector $|x, s\rangle$ with a position index x and a spin index s , a physicist says: “the particle has two quantum numbers, x and s .” When we add a colour index $a \in \{r, g, b\}$ to get $|x, s, a\rangle$, the physicist says: “we have added a new quantum number, colour.” When we specify a Fock basis vector $|\mathbf{k}, \lambda, f, a\rangle$ with momentum, helicity, flavour, and colour, the physicist says: “each particle is labelled by four quantum numbers.”

The word “quantum” is historical: these labels were first discovered through the quantization of energy levels ($n = 1, 2, 3, \dots$ for the hydrogen atom), and the discreteness of the allowed values was the revolutionary new feature—hence “quantum” number. But in modern usage, the term applies to any index on a basis vector, whether discrete ($s \in \{\uparrow, \downarrow\}$) or continuous ($\mathbf{k} \in \mathbb{R}^3$), whether arising from spatial symmetry (ℓ, m from angular momentum) or from internal symmetry (a from colour, T_3 from isospin).

This book	Physics jargon	Example
Index on a basis vector	Quantum number	n, ℓ, m, s, a
Adding an index	Adding a quantum number	$ x\rangle \rightarrow x, s\rangle$
The set of allowed values	The spectrum of the quantum number	$s \in \{\uparrow, \downarrow\}$
Direct product of spaces	Combining quantum numbers	$\mathcal{H}_x \otimes \mathcal{H}_s$
A specific basis vector	A state with definite quantum numbers	$ n = 2, \ell = 1, m = 0, s = \uparrow\rangle$
A superposition	Indefinite quantum number	$\alpha \uparrow\rangle + \beta \downarrow\rangle$: spin indefinite

The translation is mechanical: wherever a physicist says “quantum number,” read “index on a basis vector.” Wherever a physicist says “the particle has definite quantum numbers,” read “the state is a single basis vector (one-hot).” Wherever a physicist says “the quantum number is indefinite,” read “the state is a superposition (dense) in that index.”

This is why “adding a quantum number” means extending the basis vector with one more index—which is a direct product with a new Hilbert space, exactly as we constructed in the Fock section.

27.8 Labels: chirality, flavour, and the actor’s name

Having established that indices are slots acted on by matrices, we now sharpen the distinction between *indices* and *labels*. A label is a name attached to the whole field, not a slot inside it. No matrix acts along a label; no summation runs over a label. Labels are metadata that record which representation the field was declared to belong to when it was introduced into the theory.

Chirality labels L and R . The most important labels in the Standard Model are the chirality labels L and R , which we have already met. When we write ψ_L and ψ_R , the subscripts L and R are not indices. They record:

- ψ_L : this field was declared to be in the $(\frac{1}{2}, 0)$ representation of $SL(2, \mathbb{C})$. Its spinor index is undotted: ψ_L^s , $s \in \{1, 2\}$.
- ψ_R : this field was declared to be in the $(0, \frac{1}{2})$ representation. Its spinor index is dotted: $\psi_R^{\dot{s}}$, $\dot{s} \in \{1, 2\}$.

The label L or R tells you which index type the field carries. Once you see the index—dotted or undotted—the label has done its job. Both pieces of information are kept in practice because the label is easier to read at a glance, while the index is necessary for writing transformation laws and contractions.

A Lorentz transformation never converts ψ_L into ψ_R : the label is fixed at the time the field is defined, and no matrix in the theory acts on it. This is why we said that the label is a **dtype**, not a value.

Flavour labels. The Standard Model contains three generations of fermions. Within each generation, the quarks and leptons have the same gauge quantum numbers but different masses.

The generation label—called *flavour*—is another label, not an index:

$$\text{quarks: } u, d, c, s, t, b, \quad \text{leptons: } e, \mu, \tau, \nu_e, \nu_\mu, \nu_\tau. \quad (27.12)$$

We write u_R, d_L, e_R, ν_{eL} , and so on. The flavour label (the letter u, d, e , etc.) and the chirality label (L or R) together name the field completely. Neither is an index; neither is summed over.

When all three generations are treated simultaneously—for example when writing the CKM matrix that mixes quark flavours—a *flavour index* $f \in \{1, 2, 3\}$ is sometimes introduced. But at that point the flavour label becomes an index, and the CKM matrix $V^f_{f'}$ is the matrix that acts on it. The label has been promoted to an index with a transformation law. Until that promotion happens, u, d, c , etc. are just names.

What labels and indices look like side by side. To make the distinction concrete, here is the left quark doublet written first with labels only, then with all indices explicit:

Notation	Meaning
Q_L	Label only. Q : quark doublet. L : left-chiral.
Q_L^i	Internal (isospin) index $i \in \{1, 2\}$ shown; spinor suppressed.
Q_L^α	Colour index $\alpha \in \{r, g, b\}$ shown; spinor and isospin suppressed.
$Q_L^{i\alpha}$	Both gauge indices shown; spinor suppressed.
$(Q_L^{i\alpha})^s$	All indices explicit: spinor $s \in \{1, 2\}$, isospin i , colour α .

The label Q_L is present in every row. The indices grow as more slots are made explicit. In the last row, the spinor index s is written as a superscript on the whole object $(Q_L^{i\alpha})^s$, which is the clearest typographic convention: the gauge indices i, α are intrinsic to the field declaration, while the spinor index s runs over the components of the spinor representation.

Convention: which indices are superscripts and which are subscripts. Throughout this book we use the following placement:

- Undotted spinor index s : upper (superscript), transforms with $L^s_{s'}$.
- Dotted spinor index $\dot{s}$: lower (subscript), transforms with $\bar{L}^{\dot{s}}_{\dot{s}'}$ when raised, or with the conjugate when lowered. We write $\Psi_{\dot{s}}$ for the lowered dotted index.
- Internal index i, α , etc.: upper (superscript) for a field in the fundamental representation, lower for a field in the anti-fundamental. The contraction $\bar{q}_\alpha q^\alpha$ is gauge invariant precisely because one index is up and the other is down.
- Spacetime index μ : follows the standard physics convention; upper for contravariant (x^μ), lower for covariant ($x_\mu = \eta_{\mu\nu} x^\nu$).

27.9 The Dirac spinor in full index notation

We now return to the Dirac spinor and write it out with all indices explicit, in the language we have developed.

Recap: two separate fields. A left Weyl spinor $\psi_L^s(x)$ and a right Weyl spinor $\psi_R^{\dot{s}}(x)$ are two

completely independent fields:

$$\psi_L^s(x), \quad s \in \{1, 2\}, \quad \text{no dotted index}, \quad \psi_L^s \mapsto L^s_{s'} \psi_L^{s'}. \quad (27.13)$$

$$\psi_R^{\dot{s}}(x), \quad \dot{s} \in \{1, 2\}, \quad \text{no undotted index}, \quad \psi_R^{\dot{s}} \mapsto \bar{L}^{\dot{s}}_{\dot{s}'} \psi_R^{\dot{s}'}. \quad (27.14)$$

Each lives in its own $\mathbb{C}^2$: the left spinor space (undotted) and the right spinor space (dotted). These are two different $\mathbb{C}^2$'s—they carry different representations and their indices cannot be contracted with each other.

Concatenation: the Dirac spinor as a 4×1 array. The Dirac spinor Ψ is formed by *concatenating* the two 2×1 arrays into a single 4×1 array. In block form:

$$\Psi(x) = \begin{pmatrix} \psi_L(x) \\ \psi_R(x) \end{pmatrix} = \begin{pmatrix} \psi_L^1(x) \\ \psi_L^2(x) \\ \psi_R^1(x) \\ \psi_R^2(x) \end{pmatrix} \in \mathbb{C}^4. \quad (27.15)$$

The four components have a mixed index structure: the first two are indexed by $s \in \{1, 2\}$ (undotted, left), and the last two are indexed by $\dot{s} \in \{1, 2\}$ (dotted, right). The Dirac spinor is *not* a field with a single index running from 1 to 4 in a uniform sense—it is an array whose four slots belong to two different representation spaces, glued together by concatenation.

To make this precise, we introduce a single *Dirac index* $A \in \{1, 2, 3, 4\}$ as a shorthand for the combined slot:

$$\Psi^A(x) = \begin{cases} \psi_L^s(x), & A = s \in \{1, 2\} \quad (\text{upper block}), \\ \psi_R^{\dot{s}}(x), & A = \dot{s} + 2 \in \{3, 4\} \quad (\text{lower block}). \end{cases} \quad (27.16)$$

This is purely a notational convenience. The index A does not transform uniformly under $SL(2, \mathbb{C})$: the upper block transforms with L and the lower block transforms with $\bar{L}$. The transformation law in terms of A is:

$$\Psi^A(x) \mapsto S(L)^A_B \Psi^B(\Lambda^{-1}x), \quad S(L)^A_B = \begin{pmatrix} L^s_{s'} & 0 \\ 0 & \bar{L}^{\dot{s}}_{\dot{s}'} \end{pmatrix}, \quad (27.17)$$

where the 4×4 matrix $S(L)$ is block-diagonal, with L in the upper-left 2×2 block and $\bar{L}$ in the lower-right 2×2 block. The Dirac index A is thus a composite index that packages two heterogeneous spinor indices together.

Why concatenation is not the same as adding a new index. It is tempting to think of the Dirac spinor as having a “chirality index” $c \in \{L, R\}$ in addition to a spinor index $s \in \{1, 2\}$, giving a 2×2 array. This would be:

$$\text{Wrong picture: } \Psi^{c,s}(x), \quad c \in \{L, R\}, \quad s \in \{1, 2\}. \quad (27.18)$$

This is wrong for two reasons. First, $c \in \{L, R\}$ is a label, not an index—no matrix acts on it, and it cannot be summed over. Second, the undotted index s for the L block and the dotted index $\dot{s}$ for the R block are different types: they belong to different representation spaces, transform with different matrices (L vs $\bar{L}$), and cannot be combined into a single index of the same type.

The correct picture is concatenation: two separate 2×1 arrays are stacked into one 4×1 array. The chirality label L or R then serves as a *block address*—it tells you which rows of the 4×1 array belong to which representation—but it is not an index of the tensor. In PyTorch terms, this is `torch.cat([psi_L, psi_R], dim=0)`: two tensors of shape `[2]` concatenated along axis 0 to give shape `[4]`.

Adding gauge indices to the Dirac spinor. If the Dirac field also carries an internal index $i \in \{1, \dots, N\}$, the full field in explicit notation is:

$$\Psi^{A,i}(x), \quad A \in \{1, 2, 3, 4\}, \quad i \in \{1, \dots, N\}, \quad (27.19)$$

which in PyTorch has shape `[dirac=4, internal=N]`. The transformation law is:

$$\Psi^{A,i}(x) \mapsto S(L)^A{}_B U^i{}_j \Psi^{B,j}(\Lambda^{-1}x), \quad (27.20)$$

where $S(L)$ acts on the Dirac index A (mixing the four spinor components within each chirality block) and U acts on the internal index i (rotating the gauge components uniformly across all four spinor slots). The two actions commute because they act on different axes.

Expanding equation (27.20) into the upper and lower blocks:

$$\psi_L^{s,i}(x) \mapsto L^s{}_{s'} U^i{}_j \psi_L^{s',j}(\Lambda^{-1}x), \quad (27.21)$$

$$\psi_R^{\dot{s},i}(x) \mapsto \bar{L}^{\dot{s}}{}_{\dot{s}'} U^i{}_j \psi_R^{\dot{s}',j}(\Lambda^{-1}x). \quad (27.22)$$

The gauge matrix $U^i{}_j$ is identical in both lines—it does not know or care which chirality block it is acting on. The Lorentz matrices L and $\bar{L}$ are different in the two blocks, reflecting the different representations of the two chiralities.

27.10 Local gauge invariance $SU(N)$: matter demands the force

Let us set aside the spinor structure for a moment. Fix the Dirac index A (or equivalently the Weyl indices s and $\dot{s}$) and focus entirely on the internal index $i \in \{1, \dots, N\}$. The field at each spacetime point x is now simply an N -component complex vector $\psi^i(x)$, one component per internal degree of freedom. We ask: what happens when we allow a different $SU(N)$ rotation at each spacetime point?

Global $SU(N)$: no problem. A global $SU(N)$ transformation applies the same unitary matrix $U \in SU(N)$ at every spacetime point simultaneously:

$$\psi^i(x) \mapsto U^i{}_j \psi^j(x), \quad U \in SU(N), \quad \text{same } U \text{ for all } x. \quad (27.23)$$

The kinetic term $\partial_\mu \psi^i \partial^\mu \psi_i^\dagger$ is invariant under (27.23) because U is constant: it passes through the derivative, and $U^\dagger U = \mathbf{1}$ eliminates it. The mass term $m \psi^i \psi_i^\dagger$ is similarly invariant. There is no difficulty.

Local $SU(N)$: the derivative breaks invariance. Now allow U to depend on x :

$$\psi^i(x) \mapsto U^i{}_j(x) \psi^j(x), \quad U(x) \in SU(N), \quad \text{different } U \text{ at each } x. \quad (27.24)$$

The mass term $m \psi^i \psi_i^\dagger$ remains invariant: it involves ψ and $\psi^\dagger$ at the same point x , so $U(x)U^\dagger(x) = \mathbf{1}$ still cancels. But the kinetic term is no longer invariant. Under (27.24):

$$\partial_\mu \psi^i(x) \mapsto \partial_\mu (U^i{}_j(x) \psi^j(x)) = U^i{}_j(x) \partial_\mu \psi^j(x) + \underbrace{(\partial_\mu U^i{}_j(x)) \psi^j(x)}_{\text{extra term}}. \quad (27.25)$$

The extra term $(\partial_\mu U) \psi$ does not vanish: it is the price of allowing U to vary from point to point. The derivative compares the field at x and at $x + dx$; if the gauge transformation is different at the two points, the comparison is corrupted. The Lagrangian $\partial_\mu \psi^i \partial^\mu \psi_i^\dagger$ is not invariant under (27.24).

Matter demands the force: the gauge field. To restore invariance, we must introduce a new field $A_\mu(x) = A_\mu^a(x)T^a$ — the *gauge field* — that transforms under (27.24) in a specific way designed to cancel the extra term (27.25). The *covariant derivative*:

$$D_\mu \psi^i(x) = \partial_\mu \psi^i(x) - ig A_\mu^a(x) (T^a)^i_j \psi^j(x) \quad (27.26)$$

replaces ∂_μ , where T^a are the generators of $SU(N)$ and g is the coupling constant. If $A_\mu(x)$ transforms as:

$$A_\mu(x) \mapsto U(x) A_\mu(x) U^\dagger(x) + \frac{i}{g} (\partial_\mu U(x)) U^\dagger(x), \quad (27.27)$$

then $D_\mu \psi^i(x)$ transforms exactly as $\psi^i(x)$ does — covariantly, without the extra term:

$$D_\mu \psi^i(x) \mapsto U^i_j(x) D_\mu \psi^j(x). \quad (27.28)$$

The Lagrangian built with D_μ instead of ∂_μ :

$$\mathcal{L}_{\text{matter}} = \bar{\psi}^i i\gamma^\mu (D_\mu)^i_j \psi^j - m \bar{\psi}^i \psi_i \quad (27.29)$$

is now invariant under local $SU(N)$ transformations. The gauge field A_μ^a was not put in by choice; it was forced into existence by the requirement that the derivative term be locally invariant. The matter field demanded a companion force field to make local gauge invariance possible.

A common narrative to resist: symmetry does not cause force. The standard textbook story runs as follows: “we demand local gauge symmetry, and this symmetry forces the existence of the gauge field, which is the force carrier.” The photon, gluons, and $W^\pm$, Z^0 bosons are then presented as consequences of local symmetry, as if symmetry were the physical cause and the force field were the effect.

This narrative is misleading, and it obscures what is actually happening. Local gauge invariance is not a physical symmetry in the same sense as Lorentz invariance. Lorentz invariance is a true physical statement: the laws of physics are the same in all inertial frames, and this is a testable, non-trivial fact about nature. Local gauge invariance, by contrast, is a *redundancy*: field configurations related by a local gauge transformation $U(x)$ are not different physical states — they are the same physical state, described in different ways. The gauge orbit $[A_\mu] = \{A_\mu^U : U(x) \in SU(N)\}$ is the set of all field configurations that describe the same physical situation. The physical variable is not A_μ but the gauge orbit $[A_\mu]$.

The real reason to introduce the gauge field is therefore not symmetry but *dimension reduction*. The original field $\psi^i(x)$ has N complex components at each spacetime point, but many of these components are gauge-equivalent: they differ only by a local $U(x)$ rotation and carry no additional physical information. By introducing the gauge field $A_\mu^a(x)$ and demanding local gauge invariance, we are specifying that the physical degrees of freedom are the equivalence classes of (ψ^i, A_μ^a) under local $SU(N)$ transformations. This is a dramatic reduction: from the enormous space of all field configurations to the much smaller space of gauge orbits. The Faddeev-Popov procedure and BRST cohomology implement this reduction in the path integral, and the physical Hilbert space is the BRST cohomology $\ker Q / \text{im } Q$: only gauge-invariant combinations of creation operators create physical states.

Without this reduction, the Fock space would be ill-defined: the gauge field A_μ^a has four components in μ , but only two are physical (the transverse polarisations); the other two (the longitudinal and temporal polarisations) are gauge artefacts with negative or zero norm that would make the Fock space non-positive definite and the Hamiltonian unbounded below. Local gauge invariance — implemented through the covariant derivative, the gauge kinetic term, and the BRST procedure — is the structural principle that projects out these unphysical degrees of freedom and leaves a well-behaved Fock-Hamiltonian theory.

The gauge kinetic term. The gauge field itself needs a kinetic term to propagate. The *field strength tensor*:

$$F_{\mu\nu}^a(x) = \partial_\mu A_\nu^a(x) - \partial_\nu A_\mu^a(x) + g f^{abc} A_\mu^b(x) A_\nu^c(x), \quad (27.30)$$

where f^{abc} are the structure constants of $SU(N)$, transforms covariantly under (27.27). The gauge kinetic term:

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu} \quad (27.31)$$

is a Lorentz scalar (indices μ, ν contracted with $\eta^{\mu\nu}$) and a gauge scalar (index a contracted with itself, which is $SU(N)$ -invariant by the adjoint representation). It is the unique renormalizable kinetic term for the gauge field consistent with both Lorentz and gauge invariance. For $SU(3)_c$ (QCD), it contains the gluon self-interaction terms $g f^{abc} A_\mu^b A_\nu^c$ from the non-Abelian structure constants: gluons carry colour charge and interact with each other, unlike photons in the Abelian $U(1)$ case where $f^{abc} = 0$.

The complete locally gauge-invariant Lagrangian. The full Lagrangian for a matter field ψ^i coupled to a gauge field A_μ^a is:

$$\mathcal{L} = \bar{\psi}^i i\gamma^\mu (D_\mu)^i_j \psi^j - m \bar{\psi}^i \psi_i - \frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu}. \quad (27.32)$$

This is Lorentz invariant (all spacetime indices contracted with $\eta^{\mu\nu}$), locally $SU(N)$ gauge invariant (all internal indices contracted in gauge-invariant combinations), and the minimal such Lagrangian consistent with renormalizability. The coupling between matter and force is entirely contained in the covariant derivative (27.26): the single parameter g controls all interactions between the matter field ψ^i and the gauge field A_μ^a . This minimality — one coupling, one field strength, one covariant derivative — is the signature of a gauge theory, and it is why the Standard Model, despite its apparent complexity, is actually a minimal and elegant structure.

28 Multiple Internal Indices for Gauge Transformation

28.1 Multiple internal indices: product gauge groups

In the previous section we added a single internal index $i \in \{1, \dots, N\}$ to the field variable at each spacetime point x . This index is rotated by an $SU(N)$ matrix $U_j^i(x)$ that may depend on x but is blind to the spinor indices $(s, \dot{s})$ or A : it acts purely in the internal i -space, independently of the Lorentz structure. Requiring the Lagrangian to be invariant under this local $SU(N)$ transformation forced the introduction of one gauge field $A_\mu^a(x)$, which serves to make the covariant derivative well-defined on gauge orbits and reduces the physical degrees of freedom accordingly.

We can now add a second internal index $\alpha \in \{1, \dots, M\}$ to the field. This index is rotated by a separate $SU(M)$ matrix $V^\alpha_\beta(x)$ that may also depend on x , but is blind to the spinor indices $(s, \dot{s})$, blind to the first internal index i , and acts purely in the α -space. The field variable at each spacetime point is now a four-way object: spacetime location, spinor index, first internal index i , second internal index α . The two gauge transformations act independently on their respective indices and commute with each other, so the combined symmetry group is the direct product $SU(N) \times SU(M)$.

Each factor of this direct product is a separate local gauge invariance, and each requires its own gauge field to make the covariant derivative well-defined: $A_\mu^a(x)$ for the $SU(N)$ factor and $B_\mu^b(x)$ for the $SU(M)$ factor. Each gauge field independently enables a dimension reduction — from fields to gauge orbits of its respective group — and the physical Hilbert space is the intersection of both reductions. The Standard Model gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$ is precisely of this form: three independent local gauge invariances, each with its own gauge field (gluons for $SU(3)_c$, W bosons for $SU(2)_L$, the B boson for $U(1)_Y$), each performing its own dimension reduction, and together accounting for all three non-gravitational forces. The rest of this section makes this structure precise.

28.2 Adding a second internal index

In the previous section we added one internal index $i \in \{1, \dots, N\}$ to the field. We now add a second internal index, independent of the first:

$$\Psi_{\dot{s}}^{s, i, a}(x), \quad i \in \{1, \dots, N\}, \quad a \in \{1, \dots, M\}, \quad (28.1)$$

where s and $\dot{s}$ are the Lorentz spinor indices as before, i is the first internal index (acted on by $SU(N)$), and a is the second internal index (acted on by a group $SU(M)$, with matrix V). The field is now a *four-way tensor* in the ML sense, with shape `[spinor_left=2, spinor_right=2, internal_N=N, internal_M=M]`.

The second internal index in isolation. For each fixed $(i, s, \dot{s})$, the values

$$\begin{pmatrix} \Psi_{\dot{s}}^{s, i, 1}(x) \\ \Psi_{\dot{s}}^{s, i, 2}(x) \\ \vdots \\ \Psi_{\dot{s}}^{s, i, M}(x) \end{pmatrix} \in \mathbb{C}^M \quad (28.2)$$

form an $M \times 1$ column vector. A matrix $V \in SU(M)$ acts on this vector by left multiplication:

$$\Psi_{\dot{s}}^{s, i, a}(x) \longmapsto V^a_b \Psi_{\dot{s}}^{s, i, b}(x), \quad (28.3)$$

where b is contracted (summed) and a is the free output index. The matrix V^a_b acts on the a -axis for every fixed value of $(i, s, \dot{s})$ simultaneously, and acts identically on every such slice. It is completely blind to the values of i , s , and $\dot{s}$.

28.3 Both internal indices acted on simultaneously

Now consider acting with both $U \in SU(N)$ (on index i) and $V \in SU(M)$ (on index a) at the same time. For fixed $(s, \dot{s})$, the field $\Psi_{\dot{s}}^{s, i, a}$ is a 2-way array (a matrix in the ML sense) of shape $[N, M]$. The two group transformations act independently on the two axes of this matrix:

$$\Psi_{\dot{s}}^{s, i, a}(x) \longmapsto U_{i'}^i V_b^a \Psi_{\dot{s}}^{s, i', b}(x), \quad (28.4)$$

where i' and b are both contracted (summed), and i and a are both free output indices. Written out explicitly as a matrix operation on the $[N, M]$ slice:

$$\Psi_{\dot{s}}^s(x) \longmapsto U \Psi_{\dot{s}}^s(x) V^T, \quad (28.5)$$

where $\Psi_{\dot{s}}^s(x)$ is the $N \times M$ matrix with entries $(\Psi_{\dot{s}}^s)^{i, a}$, U acts by left multiplication on the row index i , and V^T acts by right multiplication on the column index a . In PyTorch:

```
einsum('ii', ab, i'b -> ia', U, V, Psi_slice).
```

The two actions commute because they act on different axes: U does not touch the a -index and V does not touch the i -index. The same pair (U, V) acts on every spinor slice $(s, \dot{s})$ identically.

28.4 Full index notation with all four indices

Restoring the Lorentz spinor indices, the complete transformation law under a simultaneous Lorentz transformation $L \in SL(2, \mathbb{C})$ and internal gauge transformations $U \in SU(N)$, $V \in SU(M)$ is:

$$\boxed{\Psi_{\dot{s}}^{s, i, a}(x) \longmapsto L_{s'}^s \bar{L}^{\dot{s}'}_{\dot{s}} U_{i'}^i V_b^a \Psi_{\dot{s}'}^{s', i', b}(\Lambda^{-1}x).} \quad (28.6)$$

There are four independent matrix actions, one for each index slot:

Index	Range	Matrix	Group	Contracted index
s	$\{1, 2\}$	$L_{s'}^s$	$SL(2, \mathbb{C})$ left	s'
$\dot{s}$	$\{1, 2\}$	$\bar{L}^{\dot{s}'}_{\dot{s}}$	$SL(2, \mathbb{C})$ right	$\dot{s}'$
i	$\{1, \dots, N\}$	$U_{i'}^i$	$SU(N)$	i'
a	$\{1, \dots, M\}$	V_b^a	$SU(M)$	b

All four matrices commute with each other in their action on Ψ because they act on different axes. Fixing any three indices and varying the fourth always gives a vector in some representation space, and the corresponding matrix acts on that vector.

28.5 Suppressing indices

Equation (28.6) is the complete, honest description of the field. In practice, various indices are suppressed depending on which structure one wants to highlight. There are two qualitatively different reasons to suppress an index.

Reason 1: the index is a singlet (size 1). If a field does not participate in $SU(M)$ —that is, it is a singlet under $SU(M)$ —then the index a takes only the value $a = 1$, the matrix V is just the number 1, and the transformation (28.3) is trivial. We drop a from the notation entirely. Similarly, if a field is a $U(1)$ singlet for the i -index, $U = e^{i\alpha}$ is a phase, and we record the charge Y as a label rather than keeping the index i explicit.

For the Lorentz spinor indices: a left Weyl spinor ψ_L^s has no dotted index (the right spinor space is absent), so $\dot{s}$ is suppressed. A right Weyl spinor $\psi_R^{\dot{s}}$ has no undotted index, so s is suppressed. A scalar field has neither.

Reason 2: the index is non-trivial but handled implicitly. The Lorentz spinor indices $(s, \dot{s})$, even when the field has a full 2×2 spinor structure, are almost always suppressed in practice. This is not because they are singlets—they are not—but because the gamma matrices γ^μ and the Dirac conjugation $\bar{\Psi} = \Psi^\dagger \gamma^0$ handle all spinor index contractions automatically. Any Lorentz-invariant expression written with gamma matrices has its spinor indices correctly contracted by the Clifford algebra, without the need to write s and s' explicitly.

When we write $\Psi^{i,a}(x)$ —suppressing $(s, \dot{s})$ but keeping both internal indices—we mean: the full field $\Psi_{\dot{s}}^{s,i,a}(x)$ with the spinor structure handled implicitly by the gamma matrices. The suppressed indices are still present and still transform; they are simply not written.

The suppression hierarchy in practice.

Notation	What is shown / suppressed
$\Psi_{\dot{s}}^{s,i,a}(x)$	Everything explicit.
$\Psi^{i,a}(x)$	Spinor indices suppressed (implicit in γ^μ).
$\Psi^i(x)$	Spinor and a -index suppressed (a : singlet or implicit).
$\Psi(x)$	All indices suppressed; structure inferred from context.

28.6 Adding labels and the Dirac spinor

Labels in the multi-index setting. As the number of indices grows, so does the importance of labels. The chirality label L or R records which spinor representation the field was declared in; it is not an index and is never summed over. Flavour labels (u, d, e, μ , etc.) name the field within a generation. These labels float alongside the field symbol regardless of how many indices are shown:

$$(Q_L)^{s,i,a}(x) \quad \text{vs} \quad (Q_L)^{i,a}(x) \quad \text{vs} \quad Q_L^{i,a}(x). \quad (28.7)$$

In all three expressions, Q_L is the label (left quark doublet), and i, a, s are indices. The label is unaffected by suppression.

The Dirac spinor with two internal indices. If we use a Dirac spinor Ψ (stacking left and right Weyl spinors) rather than a single Weyl spinor, the concatenation structure carries over unchanged. With both internal indices, the Dirac field is:

$$\Psi^{A,i,a}(x), \quad A \in \{1, 2, 3, 4\}, \quad i \in \{1, \dots, N\}, \quad a \in \{1, \dots, M\}, \quad (28.8)$$

where the Dirac index A packages the two spinor blocks:

$$\Psi^{A,i,a}(x) = \begin{pmatrix} \psi_L^{s,i,a}(x) \\ \psi_R^{\dot{s},i,a}(x) \end{pmatrix}, \quad (28.9)$$

and the full transformation law is:

$$\Psi^{A,i,a}(x) \mapsto S(L)^A_B U^i_{i'} V^a_b \Psi^{B,i',b}(\Lambda^{-1}x), \quad (28.10)$$

with $S(L) = \text{diag}(L, \bar{L})$ as before. Expanding into blocks:

$$\psi_L^{s,i,a}(x) \mapsto L^s_{s'} U^i_{i'} V^a_b \psi_L^{s',i',b}(\Lambda^{-1}x), \quad (28.11)$$

$$\psi_R^{\dot{s},i,a}(x) \mapsto \bar{L}^{\dot{s}}_{\dot{s}'} U^i_{i'} V^a_b \psi_R^{\dot{s}',i',b}(\Lambda^{-1}x). \quad (28.12)$$

The internal matrices U and V are identical in both blocks: they act the same way on every chirality. The Lorentz matrices differ between blocks, as always. This is the full four-way tensor structure, now with the spinor axis packaged into the Dirac index A .

28.7 The Standard Model gauge group $U(1)_Y \times SU(2)_L \times SU(3)_c$

The Standard Model has exactly three internal index types, corresponding to its three gauge group factors. Let us now identify N and M and add a third factor concretely.

The full internal symmetry group is:

$$G_{\text{SM}} = U(1)_Y \times SU(2)_L \times SU(3)_c, \quad (28.13)$$

and the three factors act on three independent internal index slots:

Group	Index	Range	Matrix	Physical name
$U(1)_Y$	(none; phase only)	$\{1\}$ singlet	$e^{iY\alpha}$	Hypercharge
$SU(2)_L$	i	$\{1, 2\}$	$V^i_{i'} \in SU(2)$	Weak isospin
$SU(3)_c$	α	$\{r, g, b\}$	$U^\alpha_{\alpha'} \in SU(3)$	Colour

The $U(1)_Y$ factor is special: since $N = 1$, the index is a singlet and is suppressed. The field simply acquires a phase $e^{iY\alpha}$ under $U(1)_Y$, where Y is the *hypercharge*, a rational number recorded as a label on the field.

The full field, with all indices explicit (using a Weyl spinor for concreteness), is:

$$\psi^{s,i,\alpha}(x), \quad s \in \{1, 2\}, \quad i \in \{1, 2\}, \quad \alpha \in \{r, g, b\}, \quad (28.14)$$

and it transforms under $G_{\text{SM}} \times SL(2, \mathbb{C})$ as:

$$\psi^{s,i,\alpha}(x) \mapsto e^{iY\alpha(x)} L^s_{s'} V^i_{i'} U^\alpha_{\alpha'} \psi^{s',i',\alpha'}(\Lambda^{-1}x). \quad (28.15)$$

There are four independent actions: a Lorentz action on s , an isospin action on i , a colour action on α , and a hypercharge phase on the whole field. All four commute because they act on different slots (or, in the case of $U(1)_Y$, on no slot at all—just an overall scalar multiplication).

For a Dirac spinor with full Standard Model gauge structure, the field becomes:

$$\Psi^{A,i,\alpha}(x), \quad A \in \{1, 2, 3, 4\}, \quad i \in \{1, 2\}, \quad \alpha \in \{r, g, b\}, \quad (28.16)$$

with transformation:

$$\Psi^{A,i,\alpha}(x) \mapsto e^{iY\alpha(x)} S(L)^A_B V^i_{i'} U^\alpha_{\alpha'} \Psi^{B,i',\alpha'}(\Lambda^{-1}x). \quad (28.17)$$

This is the most general fermion in the Standard Model, written with all indices honest and all labels recorded as metadata. Each Standard Model fermion is a specific instance of this template, with particular values of Y and particular representations under $SU(2)_L$ and $SU(3)_c$.

Field	Label	Y	$SU(2)_L$	$SU(3)_c$	Indices shown
Left quark	Q_L	$+\frac{1}{6}$	doublet $i \in \{1, 2\}$	triplet $\alpha \in \{r, g, b\}$	s, i, α
Right up	u_R	$+\frac{2}{3}$	singlet	triplet	s, α
Right down	d_R	$-\frac{1}{3}$	singlet	triplet	s, α
Left lepton	L_L	$-\frac{1}{2}$	doublet	singlet	s, i
Right lepton	e_R	-1	singlet	singlet	s only

Reading across each row: the label names the field and its chirality; Y is a number (not an index) recording the $U(1)_Y$ charge; the $SU(2)_L$ and $SU(3)_c$ columns tell you how many index values the corresponding axes take; and the final column lists all the indices that appear in the full explicit notation.

Singlet entries mean that index is absent: the right lepton e_R has no isospin index (it does not feel the weak force) and no colour index (it is not a quark). Its only index is the spinor index s , and it transforms only under $U(1)_Y$ (a phase) and $SL(2, \mathbb{C})$ (a Lorentz rotation).

28.8 The Beijing Opera analogy

A Standard Model fermion like the left quark doublet $Q_L^{s,i,\alpha}(x)$ carries four simultaneous layers of identity. No single layer alone identifies the particle; it is their combination that determines how the particle behaves in every interaction. This is precisely the structure of a character in a Beijing Opera performance.

In a Beijing Opera, every actor on stage carries several layers of identity simultaneously:

- The **role type** (general, scholar, clown, young warrior): a fixed category, assigned before the performance begins, that determines the broad rules of the character's behaviour.
- The **face paint** (*lian pu*): an elaborate, tradition-encoded pattern of colours and shapes, unique to each character type, applied before the actor steps on stage and unchanging throughout.
- The **costume colour and style**: determined by the role and the scene, signalling the character's allegiance, rank, and emotional state.
- The **position on stage**: where the actor stands at each moment, changing continuously throughout the performance.

No single layer is the character. The scholar-role actor in white face paint, wearing a blue robe, standing at centre stage—all four layers together make the character.

The correspondence with a Standard Model field is exact:

Beijing Opera	Field theory	Technical meaning
Position on stage	Spacetime point x	Where the field is evaluated; continuous, always suppressed in notation.
Face paint (<i>lian pu</i>)	Lorentz spinor index s or $\dot{s}$	How the field transforms under rotations and boosts; acted on by L or $\bar{L}$.
Costume colour	Colour index $\alpha \in \{r, g, b\}$	How the field transforms under the strong force; acted on by $U \in SU(3)_c$.
Role type	Isospin index $i \in \{1, 2\}$	Whether the field participates in the weak force as a doublet partner; acted on by $V \in SU(2)_L$.
Actor's name	Chirality and flavour labels ($L, R, u, d, e, \dots$)	The field's identity; not an index, never acted on by any matrix.
Hypercharge	Y (a number, not an index)	Overall phase under $U(1)_Y$; a property of the whole field, like the actor's seniority rank.

The left quark doublet $Q_L^{s,i,\alpha}(x)$ is: the actor named Q_L (label), wearing the spinor face paint (index s), in the coloured quark costume (index α), playing the doublet role (index i), standing at position x on the stage. All layers together make the character; remove any one layer and the identity is incomplete.

There is one further parallel that the Beijing Opera makes vivid: the *independence of the layers*. The costume department does not consult the face painter; the stage manager does not consult the costume department. Each layer is controlled by its own rule, and the rules do not interfere. This is precisely what we mean when we say that the four matrices L , V , U , $e^{iY\alpha}$ commute in their action on $\Psi^{s,i,\alpha}(x)$: they act on different index slots and do not talk to each other. The strong force rotates the colour index α without touching s or i ; the weak force rotates the isospin index i without touching s or α ; the Lorentz transformation rotates the spinor index s without touching i or α . The drama on stage is the combination of all four simultaneous actions.

Remark. [ML analogy: multi-head, multi-channel tensors] In a transformer model, the hidden state at each position carries multiple simultaneous index types: a head index, a feature index, a sequence-position index. Each index is acted on by its own set of weight matrices, and the different operations (attention over heads, feed-forward over features, positional encoding) are independent of each other. A Standard Model field is the same structure, with the symmetry group playing the role of the weight matrices. The Beijing Opera analogy and the transformer analogy are the same analogy, seen from different cultural vantage points.

28.9 Nature is left-handed: isospin as a chiral index

Look again at the table of Standard Model fermions from the previous section. Focus on the $SU(2)_L$ column:

Field	Label	$SU(2)_L$ rep	Isospin index i
Left quark	Q_L	doublet	$i \in \{1, 2\}$ ✓
Right up	u_R	singlet	absent
Right down	d_R	singlet	absent
Left lepton	L_L	doublet	$i \in \{1, 2\}$ ✓
Right lepton	e_R	singlet	absent

The pattern is unambiguous: the isospin index $i \in \{1, 2\}$ appears *only* on left-chiral fields. Every right-chiral field is an $SU(2)_L$ singlet. This is not a notational choice or a convention. It is a measured, experimental fact about the universe, confirmed to extraordinary precision by collider experiments: *the weak force couples only to left-chiral fermions*.

This single observation—the presence or absence of the index i —is the precise technical statement of the slogan “Nature is left-handed.”

What it means in index language. The weak gauge bosons $W^\pm$ and Z are the gauge fields of $SU(2)_L$. A gauge coupling is, at its core, an interaction vertex where the gauge boson contracts its internal index with the internal index of the fermion it couples to. For the $W^\pm$ boson, this coupling has the schematic form:

$$g W_\mu^{+i}{}_{i'} \bar{\psi}^i(x) \gamma^\mu \psi^{i'}(x), \quad (28.18)$$

where i and i' are isospin indices contracted between the W boson and the fermion field. This

vertex exists *only if the fermion field carries an isospin index i* . If the field has no isospin index (a singlet), the contraction is impossible and the coupling does not exist.

Since only left-chiral fields carry the index i , only left-chiral fields can participate in this contraction. Right-chiral fields have no i index to contract with. The W boson simply does not see them.

The subscript L in $SU(2)_L$ is not decoration. The subscript L in the gauge group name $SU(2)_L$ is a reminder of exactly this fact. It is the $SU(2)$ that acts *only on left-chiral fields*. By contrast, $SU(3)_c$ (colour) and $U(1)_Y$ (hypercharge) couple to both chiralities, though with different hypercharge values. The asymmetry is entirely in the $SU(2)_L$ factor.

Contrast with the other gauge groups.

Group	Couples to	Chirality dependence	Index carried by
$U(1)_Y$	Both ψ_L and ψ_R	Different Y for each	Phase only (no index)
$SU(2)_L$	ψ_L only	Couples to L , blind to R	$i \in \{1, 2\}$ on L fields only
$SU(3)_c$	Both ψ_L and ψ_R	Same coupling for each	$\alpha \in \{r, g, b\}$ on both

The strong force ($SU(3)_c$) is chirality-blind: both ψ_L and ψ_R carry the colour index α , and the gluon couples to both in exactly the same way. The electromagnetic force (which emerges after symmetry breaking from $U(1)_Y$ and $SU(2)_L$ together) is also chirality-blind at low energies. Only the weak force breaks the left-right symmetry, and it does so maximally: the right-chiral fields simply have no i -index to couple to the W boson.

Why is this surprising? Before the 1950s, physicists assumed that the laws of physics were symmetric under spatial reflection (parity P), which swaps left-handed and right-handed particles. In our language, parity would exchange $\psi_L^s \leftrightarrow \psi_R^s$ and therefore require that any field carrying the isospin index i under $SU(2)_L$ must have a parity partner with the same index. If this were true, both chiralities would carry i , and the table would have doublets in every row.

The 1956 experiments of Wu et al. on cobalt-60 beta decay, and the theoretical prediction by Lee and Yang, showed that this is not the case. The weak interaction *violates parity maximally*: it couples only to left-chiral fermions. In index language: i is carried by L -fields and absent from R -fields. Nature chose an asymmetric index structure.

The Beijing Opera rephrasing. In Beijing Opera terms: the weak force is a stage manager who only interacts with actors playing the “doublet role” ($i \in \{1, 2\}$). Actors in singlet roles (R -fields) are invisible to this stage manager — he walks past them without instructions. The strong force and electromagnetism are stage managers who interact with all actors equally, regardless of role. The drama of the weak interaction is a drama of selective attention: it sees only half the cast.

After symmetry breaking: mass and the mixing of chiralities. We noted in the Dirac spinor section that a mass term couples ψ_L and ψ_R together. In the Standard Model this coupling comes from the Higgs field ϕ^i , which also carries an isospin index i . The Yukawa coupling:

$$y \bar{L}_L^i \phi_i e_R + \text{h.c.}, \quad (28.19)$$

contracts the isospin index i of the left lepton doublet L_L^i with the isospin index of the Higgs ϕ_i , and then couples the result to the right lepton singlet e_R (which has no isospin index). When the Higgs acquires a vacuum expectation value $\langle \phi \rangle$, the index i is “used up” by the contraction,

and what remains is an effective mass term $m \bar{\psi}_L \psi_R$ coupling the two chiralities. The isospin index is the bridge that connects the left-handed world to the right-handed world, via the Higgs.

Here h.c. denotes the Hermitian conjugate of the preceding term, which is $y^* \bar{e}_R \phi^{i*} L_{L,i}$; it is added to ensure the Lagrangian density is Hermitian ($\mathcal{L}^\dagger = \mathcal{L}$), as required for real energy eigenvalues and conservation of probability.

This is the deepest sense in which Nature is left-handed: not only does the weak force couple only to L -fields, but the very mechanism by which particles acquire mass — the Yukawa coupling to the Higgs — requires the isospin index to exist on the L -field in the first place. Without the asymmetric index structure, there would be no Yukawa coupling, and therefore no fermion masses. The left-handedness of the weak interaction and the origin of mass are, at the level of indices, the same story.

28.10 Nature prefers minimalism over symmetry

The previous subsection made a striking observation: the isospin index $i \in \{1, 2\}$ appears on left-chiral fields and is absent from right-chiral fields. The weak force couples only to the left. This is maximally asymmetric, and it prompts a deeper question: why would Nature choose an asymmetric theory?

The answer, we argue, is that Nature does not choose symmetry. Nature chooses *minimalism*. Symmetry is a tool — a mechanism for achieving minimalism — but it is not the organizing principle. Where symmetry serves minimalism, Nature uses it. Where symmetry would impose redundancy or unnecessary structure, Nature discards it.

The human body: a minimal design, not a symmetric one

The human body is the clearest illustration of this distinction. Viewed from the outside, the body is approximately bilaterally symmetric: two arms, two legs, two eyes, two ears, symmetrically arranged. One might conclude that symmetry is a principle of biological design.

But look inside. The heart is on the left. The liver is on the right. The left lung has two lobes; the right lung has three. The stomach, the spleen, the appendix — all asymmetrically placed. The brain's two hemispheres are anatomically similar but functionally asymmetric: language processing is predominantly left-lateralised; spatial reasoning is right-lateralised. Even at the molecular level, biological amino acids are almost exclusively left-handed (L-amino acids), despite no fundamental physical reason to prefer one chirality over the other.

The pattern is clear: *external symmetry where it serves function (two eyes for depth perception, two hands for manipulation), internal asymmetry where it serves efficiency*. A body with two hearts, two livers, two stomachs, perfectly mirrored, would be heavier, more metabolically expensive, and no more functional. The asymmetry of internal organs is not a flaw in the design — it is the design.

The Standard Model as a minimal design

The Standard Model has the same structure. It has substantial symmetry — three gauge groups, Lorentz invariance, CPT invariance — but these symmetries exist not as ends in themselves but as the minimal structure needed to produce a consistent, renormalizable, predictive theory with the observed particle content. Where symmetry would impose unnecessary structure, the Standard Model discards it.

The weak force does not need to be left-right symmetric. Parity symmetry (exchanging left and right, $\psi_L \leftrightarrow \psi_R$) would require both chiralities to carry the isospin index i . But if both

ψ_L and ψ_R carried isospin doublet structure, the weak force would couple equally to both — and a Dirac mass term $m\bar{\psi}_L\psi_R$ would be gauge invariant without the Higgs, since both sides carry the same representation. The theory would have fewer parameters, but it would be a different theory, describing a different universe. The observed universe has massive fermions generated by the Higgs mechanism, and a weak force that violates parity maximally. The asymmetric index structure — i on L , absent on R — is the minimal structure that produces this.

Gauge symmetry is a tool for consistency, not an aesthetic choice. The three gauge groups $SU(3)_c \times SU(2)_L \times U(1)_Y$ are not present because Nature has a preference for symmetry. They are present because:

- Without $SU(3)_c$, quarks would not be confined and the proton would not be stable.
- Without $SU(2)_L$, there is no mechanism that produces the observed W and Z masses, beta decay, and the weak interaction.
- Without $U(1)_Y$, the photon and the Z cannot emerge from the same sector via the Weinberg mixing.

Each gauge group is the *minimal* group required to produce a specific observed phenomenon. The gauge symmetry is the scaffolding; the phenomenon is the goal. And the scaffolding is not symmetric: three different groups, three different dimensions (1, 3, 8), three different coupling constants.

The Higgs sector is not symmetric, and it does not need to be. The Higgs doublet ϕ^i has hypercharge $Y = +\frac{1}{2}$. A perfectly symmetric electroweak sector would suggest equal treatment of ϕ and ϕ^* , equal masses for $W^\pm$ and Z^0 , no mixing angle. But the Weinberg angle θ_W is nonzero; $m_W \neq m_Z$; the photon is massless while the Z is not. The pattern of symmetry breaking is not the maximally symmetric one — it is the minimally structured one consistent with leaving electromagnetism unbroken and giving masses to the weak force carriers.

Symmetry as a tool: what it achieves

If symmetry is not the principle, what is its role? It is a *tool* for achieving two specific goals:

Goal 1: Renormalizability. A theory with arbitrary couplings between arbitrary fields is not predictive: it requires infinitely many parameters. Gauge symmetry restricts which couplings are allowed, reducing the parameter count to a finite number. The 19 parameters of the Standard Model are the minimal set needed to specify all non-gravitational physics — because gauge symmetry has eliminated all other couplings. Symmetry here is not an aesthetic principle but a *complexity reduction*: it is the theory’s way of saying “these couplings are related, not independent.”

Goal 2: Stability. A theory without gauge symmetry for its vector fields would have unphysical (ghost) polarisation modes propagating in loops, destabilising the vacuum and making the Hamiltonian unbounded below. Gauge symmetry removes these modes from the physical spectrum, as we saw in the BRST discussion. The symmetry is the mechanism for keeping the dynamics stable. It is not a statement about the elegance of the theory — it is a statement about the consistency of the theory.

Where symmetry is absent, it is absent for a reason. Parity is violated because the weak force needs to distinguish chiralities to implement the Higgs mass generation mechanism. CP is violated (weakly, via the CKM phase) because the universe contains more matter than antimatter and the Standard Model must accommodate this. Flavour symmetry between the three generations is explicitly broken by the Yukawa couplings because the quarks and leptons

have very different masses. In each case, the broken symmetry is not an imperfection — it is a minimal departure from symmetry, precisely calibrated to produce the observed physics.

Minimalism as the actual principle

What does minimalism mean, precisely?

Minimalism in field content. The Standard Model uses the minimal number of fields needed to account for the observed particles. There are no redundant fields. Every field plays a specific role: quarks and leptons are matter; gauge bosons are force carriers; the Higgs generates masses. A theory with more fields would require more parameters and would predict new particles not observed.

Minimalism in coupling structure. Every coupling in the Standard Model Lagrangian is required by either gauge invariance or renormalizability. No coupling is present “because it is symmetric.” The Yukawa coupling $y_e \bar{L}_L^i \phi_i e_R$ exists because it is the *unique* gauge-invariant, renormalizable coupling between the Higgs, the left lepton doublet, and the right electron singlet. Not because it is elegant. Because it is the only thing that works.

Minimalism in symmetry breaking. The Higgs potential $V(\phi) = -\mu^2 |\phi|^2 + \lambda |\phi|^4$ breaks electroweak symmetry *minimally*: it preserves electromagnetism (the unbroken $U(1)_{\text{em}}$ generated by $Q = T^3 + Y$) while giving masses to the $W^\pm$ and Z^0 . A more general potential could break more symmetries or break them differently. The observed pattern of symmetry breaking — three massive weak bosons, one massless photon — is the minimal departure from the symmetric phase consistent with the observed low-energy physics.

The index structure is the signature of minimalism. In the language of this book, minimalism appears as *index economy*: a field carries only the indices it needs to participate in the interactions it participates in, and no more. The right-handed electron e_R has no colour index (it does not feel the strong force), no isospin index (it does not feel the weak force), and only a single spinor index (it is a Weyl spinor, not a Dirac spinor). Its full index set is $\{s, \pm\}$ (spinor and particle/antiparticle). This is the minimum necessary to describe a charged, massive, spin- $\frac{1}{2}$ particle that feels electromagnetism and gravity only.

Field	Spinor	Colour	Isospin	Y	Why these indices?
$Q_L^{s,i,\alpha}$	$s \in \{1, 2\}$	$\alpha \in \{r, g, b\}$	$i \in \{1, 2\}$	$+\frac{1}{6}$	Needs all: matter, colour, weak
$u_R^{s,\alpha}$	$s \in \{1, 2\}$	$\alpha \in \{r, g, b\}$	absent	$+\frac{2}{3}$	No weak; isospin absent
e_R^s	$s \in \{1, 2\}$	absent	absent	-1	Only EM; colour and isospin absent
ϕ^i	absent	absent	$i \in \{1, 2\}$	$+\frac{1}{2}$	Scalar; needs isospin for Yukawa
$G^{a,\mu}$	$\mu \in \{0, \dots, 3\}$	$a \in \{1, \dots, 8\}$	absent	0	Gauge boson; colour-adjoint
B_μ	$\mu \in \{0, \dots, 3\}$	absent	absent	0	Gauge boson; no colour or isospin

Every absent index in this table is a statement about minimalism. The right-handed electron has no colour index because it does not need one. The B_μ boson has no colour or isospin index

because it is the gauge boson of $U(1)_Y$ alone. Each field carries exactly the indices required by its role in the theory, and no more.

The deeper analogy: human body and Standard Model

Human body	Standard Model
Bilateral symmetry externally (two eyes, two hands)	Lorentz and CPT symmetry: all fields transform consistently under spacetime symmetries
Heart on the left, liver on the right	$SU(2)_L$ acts on left chirality only; right-chiral fields are singlets
Left lung has 2 lobes, right has 3	e_R has no isospin index; L_L has isospin index $i \in \{1, 2\}$
L-amino acids only, despite no physical preference	CP violation via CKM phase: small but present, with observable consequences
Brain hemispheres: same anatomy, different function	ψ_L and ψ_R have same spinor dimension but different gauge representations
Efficient internal asymmetry reduces metabolic cost	Chiral gauge theory reduces parameter count and prevents redundant couplings
The body works, despite being asymmetric	The Standard Model is renormalizable and consistent, despite being chiral

In both cases, the design is not “maximally symmetric” — it is *minimal*. The bilateral symmetry of the human body is useful where it serves function (locomotion, perception, manipulation) and absent where it would impose cost without benefit (two hearts do not give twice the circulation; they give twice the failure risk). The gauge symmetry of the Standard Model is present where it ensures consistency and reduces parameters, and absent (or broken) where it would prevent the observed mass spectrum or force unnecessary equalities between couplings that are measured to be unequal.

The philosophical lesson. A common misconception about fundamental physics is that it tends toward greater and greater symmetry — that the most fundamental theory would be maximally symmetric. The Standard Model argues against this. Its gauge symmetry is broken (by the Higgs mechanism). Its discrete symmetries are broken (parity by the weak force; CP by the CKM phase). Its approximate symmetries (flavour, custodial $SU(2)$) are broken by the Yukawa couplings and the hypercharge assignments. What remains unbroken is not the most symmetry possible, but the *least symmetry necessary*: the symmetries required for consistency, renormalizability, and the correct vacuum structure.

Symmetry is the wrong question. The right question is: what is the *minimal* structure that produces the observed phenomena? The Standard Model is the answer currently known, and its asymmetries — the absent indices, the chiral couplings, the broken parity — are not imperfections to be explained away. They are the signatures of a design that is not symmetric, but is minimal.

Remark. [ML analogy: inductive bias and the absence of symmetry] In machine learning, the choice of architecture imposes an *inductive bias*: the model is not symmetric in all directions of its parameter space, because different parameters play different roles. A convolutional network is translation-equivariant but not rotation-equivariant; a transformer is permutation-equivariant across tokens but not across feature dimensions. These asymmetries are not flaws — they are design choices that reflect the structure of the task. Imposing full

permutation symmetry on all dimensions would reduce the model's expressiveness without improving its performance.

The Standard Model has the same structure. The “architecture” — the choice of gauge group, field representations, and Yukawa couplings — is not the most symmetric possible, but it is the one matched to the task: producing a consistent, renormalizable description of the observed particle spectrum. The absent symmetries are the model's inductive biases, encoding what we know about which fields interact with which forces. The right question is not “why is there so much symmetry?” but “why is there exactly this much symmetry, and no more?”

29 Fields, Lagrangian, and Noether Currents

29.1 Three kinds of field: a reminder

Before designing a Lagrangian, we collect the three species of field that appear in the Standard Model and recall their index structures.

Scalar field. A scalar field $\phi(x)$ carries no Lorentz indices. It is a single complex (or real) number at each spacetime point. Under a Lorentz transformation Λ :

$$\phi(x) \longmapsto \phi(\Lambda^{-1}x). \quad (29.1)$$

The value does not rotate; only the evaluation point moves. This is the $(0,0)$ representation. The Higgs field is a scalar (with an additional isospin index $i \in \{1,2\}$, but no spinor index).

Spinor field. A spinor field $\psi^s(x)$ (left Weyl) or $\psi^{\dot{s}}(x)$ (right Weyl) carries one spinor index and transforms as:

$$\psi^s(x) \longmapsto L^s_{s'} \psi^{s'}(\Lambda^{-1}x), \quad \psi^{\dot{s}}(x) \longmapsto \bar{L}^{\dot{s}}_{\dot{s}'} \psi^{\dot{s}'}(\Lambda^{-1}x). \quad (29.2)$$

A Dirac spinor Ψ^A concatenates both into a 4×1 column, transforming with the block-diagonal $S(L) = \text{diag}(L, \bar{L})$. These are the $(\frac{1}{2}, 0)$, $(0, \frac{1}{2})$, and their direct sum. Quarks, leptons, and neutrinos are spinor fields.

Vector field (in the physics sense). A vector field $A^\mu(x)$ carries one contravariant Lorentz index $\mu \in \{0, 1, 2, 3\}$ and transforms as:

$$A^\mu(x) \longmapsto \Lambda^\mu_{\nu} A^\nu(\Lambda^{-1}x). \quad (29.3)$$

This is the $(\frac{1}{2}, \frac{1}{2})$ representation, which corresponds to the 2×2 matrix field $A = A_\mu \sigma^\mu$ transforming as $A \mapsto L A L^\dagger$. The index μ here is a *physics* index: it transforms with the 4×4 Lorentz matrix Λ^μ_{ν} , not with the 2×2 spinor matrix L . The photon A^μ , the $W^\pm$ bosons $W^{i,\mu}$, the Z boson, and the gluons $G^{a,\mu}$ are all vector fields.

The cast of the Standard Model.

Type	Lorentz rep	Indices	Examples
Scalar	$(0,0)$	none (plus possible internal i)	Higgs ϕ^i
Spinor	$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	s or $\dot{s}$ (or Dirac A)	quarks, leptons
Vector	$(\frac{1}{2}, \frac{1}{2})$	$\mu \in \{0, 1, 2, 3\}$	γ , $W^\pm$, Z , gluons

29.2 Designing a Lagrangian: the rules of the game

The Lagrangian density $\mathcal{L}(x)$ is the central object of a quantum field theory. The action is $S = \int d^4x \mathcal{L}(x)$, and the equations of motion follow from requiring $\delta S = 0$. Quantum amplitudes are computed from e^{iS} . Every physical prediction of the Standard Model follows from one $\mathcal{L}$.

The rules for a valid Lagrangian are:

1. **It must be a Lorentz scalar.** $\mathcal{L}(x)$ must have no free Lorentz indices of any kind: no free μ , no free s , no free $\dot{s}$. Every index must be contracted.
2. **It must be real** (or more precisely, Hermitian as an operator): $\mathcal{L}^\dagger = \mathcal{L}$. This ensures the action S is real and probabilities are conserved.

3. **It must be renormalizable** (for the Standard Model): no terms with mass dimension greater than 4 in natural units. This limits the polynomial degree of the fields.
4. **It must be invariant under all gauge symmetries.** This is the deepest constraint and the subject of this section.

Rule 1 is the index-contraction rule we have been building toward. Rules 2–4 further restrict which index contractions are allowed.

ML analogy. Designing a Lagrangian is like designing a loss function for a neural network: you write down the most general expression consistent with your symmetry requirements, and the dynamics (equations of motion, Feynman rules, scattering amplitudes) emerge automatically. The symmetry constraints play the role of architectural constraints (equivariance, weight-tying) in geometric deep learning.

29.3 Kinetic and potential terms

Every Lagrangian is built from two kinds of terms.

Kinetic terms describe how fields propagate through spacetime. They involve derivatives ∂_μ and must contract all Lorentz indices to form a scalar.

For a *scalar field*:

$$\mathcal{L}_{\text{kin}}^{\text{scalar}} = \partial^\mu \phi^* \partial_\mu \phi = \eta^{\mu\nu} (\partial_\mu \phi^*) (\partial_\nu \phi). \quad (29.4)$$

The index μ on ∂^μ and the index μ on ∂_μ are contracted with the Minkowski metric $\eta^{\mu\nu}$, giving a Lorentz scalar. This term has mass dimension 4 (each ∂_μ contributes dimension 1, each ϕ contributes dimension 1 for a scalar in 4d), so it is renormalizable.

For a *Dirac spinor field*:

$$\mathcal{L}_{\text{kin}}^{\text{Dirac}} = \bar{\Psi} i \gamma^\mu \partial_\mu \Psi = i \bar{\Psi}^A (\gamma^\mu)^A_B \partial_\mu \Psi^B. \quad (29.5)$$

Here $\bar{\Psi} = \Psi^\dagger \gamma^0$. Two index contractions happen simultaneously:

- The Lorentz vector index μ on γ^μ is contracted with the μ on ∂_μ , handled by the implicit $\eta^{\mu\nu}$ in the notation $\gamma^\mu \partial_\mu$.
- The Dirac index A is contracted between $\bar{\Psi}^A$ and $(\gamma^\mu)^A_B$ and Ψ^B , summing over all four spinor components.

The result is a Lorentz scalar. The gamma matrices are the device that allows a spinor bilinear to contract with a vector derivative and produce a scalar.

For a *vector field*:

$$\mathcal{L}_{\text{kin}}^{\text{vector}} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu}, \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \quad (29.6)$$

The *field strength tensor* $F_{\mu\nu}$ is antisymmetric in μ, ν and carries two Lorentz vector indices. Contracting $F^{\mu\nu} F_{\mu\nu} = \eta^{\mu\rho} \eta^{\nu\sigma} F_{\mu\nu} F_{\rho\sigma}$ uses both metrics and contracts all four free indices, giving a scalar. This term encodes the kinetic energy of the photon (for $U(1)$) or the gauge bosons generally.

Potential terms involve no derivatives and describe self-interactions and masses. The rules are the same: all indices must be contracted.

For a scalar:

$$\mathcal{L}_{\text{pot}}^{\text{scalar}} = -V(\phi^* \phi) = -\mu^2 \phi^* \phi - \lambda (\phi^* \phi)^2. \quad (29.7)$$

The combination $\phi^* \phi$ is manifestly a Lorentz scalar (no indices) and is the only renormalizable, gauge-invariant polynomial for a charged scalar. The coefficient $-\mu^2$ is the mass-squared term; λ

controls self-interaction. When $\mu^2 < 0$, the potential has a “Mexican hat” shape and spontaneous symmetry breaking occurs—this is the Higgs mechanism.

For a Dirac spinor, a mass term requires coupling ψ_L and ψ_R :

$$\mathcal{L}_{\text{mass}}^{\text{Dirac}} = -m \bar{\Psi}\Psi = -m (\bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L). \quad (29.8)$$

In index language: $\bar{\Psi}_A \Psi^A = \sum_A \bar{\Psi}_A \Psi^A$ contracts the Dirac index. In block form, $\bar{\psi}_L \psi_R$ pairs an undotted index (from $\bar{\psi}_L$, which after conjugation carries a dotted index) with a dotted index (from ψ_R), using the epsilon tensor to form an invariant. The mass term mixes chiralities.

For a vector field, a naïve mass term $\frac{1}{2}m^2 A^\mu A_\mu$ is in principle possible but is forbidden by gauge invariance, as we shall see immediately below.

29.4 Noether’s theorem: symmetry dictates conservation

Every continuous symmetry of the Lagrangian produces a conserved current. This is Noether’s theorem, and it is one of the deepest results in theoretical physics. It connects the abstract language of symmetry groups to the concrete, measurable quantities that are conserved in experiments: energy, momentum, angular momentum, electric charge, colour charge. We derive it here in full generality, then apply it to each symmetry we have encountered.

The general derivation

Consider a Lagrangian density $\mathcal{L}(\Phi, \partial_\mu \Phi)$ depending on a field $\Phi(x)$ and its first derivatives. The equations of motion are the Euler–Lagrange equations:

$$\frac{\partial \mathcal{L}}{\partial \Phi} - \partial_\mu \frac{\partial \mathcal{L}}{\partial (\partial_\mu \Phi)} = 0. \quad (29.9)$$

Now suppose the Lagrangian is invariant under a continuous transformation parametrised by a small constant ϵ :

$$\Phi(x) \mapsto \Phi(x) + \epsilon \delta\Phi(x), \quad (29.10)$$

where $\delta\Phi$ is the infinitesimal change in the field. “Invariant” means the Lagrangian changes by at most a total derivative under (29.10):

$$\mathcal{L} \mapsto \mathcal{L} + \epsilon \partial_\mu K^\mu \quad (29.11)$$

for some K^μ (often $K^\mu = 0$). Computing the variation of $\mathcal{L}$ directly:

$$\begin{aligned} \epsilon \delta \mathcal{L} &= \frac{\partial \mathcal{L}}{\partial \Phi} \epsilon \delta\Phi + \frac{\partial \mathcal{L}}{\partial (\partial_\mu \Phi)} \epsilon \partial_\mu (\delta\Phi) \\ &= \epsilon \left[\frac{\partial \mathcal{L}}{\partial \Phi} - \partial_\mu \frac{\partial \mathcal{L}}{\partial (\partial_\mu \Phi)} \right] \delta\Phi + \epsilon \partial_\mu \left[\frac{\partial \mathcal{L}}{\partial (\partial_\mu \Phi)} \delta\Phi \right]. \end{aligned} \quad (29.12)$$

On shell (when the equations of motion (29.9) hold), the first bracket vanishes. Setting the remainder equal to $\epsilon \partial_\mu K^\mu$:

$$\partial_\mu \left[\frac{\partial \mathcal{L}}{\partial (\partial_\mu \Phi)} \delta\Phi - K^\mu \right] = 0. \quad (29.13)$$

This is a *conservation law*. The conserved current is:

$$\boxed{j^\mu = \frac{\partial \mathcal{L}}{\partial (\partial_\mu \Phi)} \delta\Phi - K^\mu, \quad \partial_\mu j^\mu = 0.} \quad (29.14)$$

The conserved *charge* is the spatial integral of the time component:

$$Q = \int d^3x j^0(x), \quad \frac{dQ}{dt} = 0. \quad (29.15)$$

The charge Q is a number that does not change with time, no matter what the field does, as long as the equations of motion hold.

ML analogy. Noether's theorem is the physics analogue of the following observation in ML: if a loss function is invariant under a group of transformations acting on the parameters, then the gradient flow (the equations of motion for gradient descent) conserves a corresponding quantity along every trajectory. The conserved current j^μ is the “Noether gradient” associated with the symmetry direction. In both cases, symmetry of the objective implies a structural constraint on the dynamics.

Spacetime symmetries: energy and momentum

The simplest application is to spacetime translation symmetry. The transformation is:

$$x^\mu \mapsto x^\mu + \epsilon^\mu, \quad \Phi(x) \mapsto \Phi(x + \epsilon) = \Phi(x) + \epsilon^\nu \partial_\nu \Phi(x), \quad (29.16)$$

so $\delta\Phi = \epsilon^\nu \partial_\nu \Phi$. Any Lagrangian with no explicit x -dependence satisfies $\delta\mathcal{L} = \epsilon^\nu \partial_\nu \mathcal{L} = \epsilon^\nu \partial_\mu (\delta_\nu^\mu \mathcal{L})$, so $K^\mu = \delta^\mu_\nu \mathcal{L}$. The Noether current for translation in the ν direction is:

$$T^\mu{}_\nu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \Phi)} \partial_\nu \Phi - \delta^\mu_\nu \mathcal{L}, \quad (29.17)$$

the *stress-energy tensor*. Conservation $\partial_\mu T^\mu{}_\nu = 0$ gives four conserved charges: energy ($\nu = 0$) and three-momentum ($\nu = 1, 2, 3$):

$$E = \int d^3x T^{00}, \quad P^k = \int d^3x T^{0k}. \quad (29.18)$$

Similarly, Lorentz rotational symmetry gives rise to the conserved angular momentum tensor, whose spatial components are the three components of angular momentum $\mathbf{J}$. *Energy, momentum, and angular momentum are all Noether charges of spacetime symmetry.*

Internal symmetries: conserved charges

Now consider an *internal* symmetry, where the field transforms but the spacetime point does not.

Global $U(1)$: electric charge. For a complex scalar with Lagrangian (??), the global $U(1)$ transformation is $\phi \mapsto e^{i\alpha} \phi$, so $\delta\phi = i\phi$ (and $\delta\phi^* = -i\phi^*$). There is no total-derivative term ($K^\mu = 0$). Applying (29.14):

$$j^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} (i\phi) + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi^*)} (-i\phi^*) = i(\phi^* \partial^\mu \phi - \phi \partial^\mu \phi^*). \quad (29.19)$$

The conserved charge is:

$$Q = i \int d^3x (\phi^* \dot{\phi} - \phi \dot{\phi}^*), \quad (29.20)$$

which, after canonical quantization, counts particles minus antiparticles. This is *electric charge* (up to normalization). The conservation law $\partial_\mu j^\mu = 0$ is the field-theory statement of charge conservation.

The two-term expression appears frequently enough to warrant a shorthand. Define the *antisymmetrised derivative*:

$$A \overleftrightarrow{\partial}_\mu B \equiv A (\partial_\mu B) - (\partial_\mu A) B, \quad (29.21)$$

so that $j^\mu = i \phi^* \overleftrightarrow{\partial}^\mu \phi$. The arrow indicates that ∂_μ acts in both directions, with a relative minus sign. The antisymmetrisation is what makes j^μ divergence-free: $\partial_\mu j^\mu = i[\phi^* \square \phi - (\square \phi^*) \phi] = 0$ when both ϕ and ϕ^* satisfy the Klein–Gordon equation. (This is the same current that appeared in eq. (13.65), where its sign-indefiniteness ruled out a single-particle probability interpretation.)

For the Dirac field Ψ , the same $U(1)$ symmetry $\Psi \mapsto e^{i\alpha} \Psi$ gives:

$$j^\mu = \bar{\Psi} \gamma^\mu \Psi = \bar{\Psi}^A (\gamma^\mu)_A^B \Psi^B, \quad (29.22)$$

the *Dirac current*. In index language: the Dirac index A is contracted between $\bar{\Psi}$ and $\gamma^\mu \Psi$, and the result is a Lorentz 4-vector (one free index μ) whose divergence vanishes on shell. This is the electromagnetic current of a fermion.

Global $SU(2)_L$: weak isospin charge. For a left-chiral doublet Ψ_L^i with global $SU(2)$ symmetry, the infinitesimal transformation is $\delta \Psi_L^i = i\theta^a (\tau^a/2)^i_{i'} \Psi_L^{i'}$ for constant parameters θ^a . There are three independent parameters $\theta^1, \theta^2, \theta^3$, one per generator, giving three independent Noether currents:

$$(j^a)^\mu = \bar{\Psi}_L^i \gamma^\mu \left(\frac{\tau^a}{2}\right)^i_{i'} \Psi_L^{i'}, \quad a = 1, 2, 3. \quad (29.23)$$

The three conserved charges are the *weak isospin charges* T^1, T^2, T^3 . In index language: both the isospin index i (via the generator $(\tau^a/2)^i_{i'}$) and the spinor index (via γ^μ) are contracted, and the result is a vector current for each generator direction a .

Global $SU(3)_c$: colour charge. For a quark field q^α with global $SU(3)$ colour symmetry, there are eight generators $T^a = \lambda^a/2$, giving eight currents:

$$(j^a)^\mu = \bar{q}_\alpha \gamma^\mu (T^a)^\alpha_{\alpha'} q^{\alpha'}, \quad a = 1, \dots, 8. \quad (29.24)$$

The eight conserved charges are the *colour charges*. A quark carries colour charge; a gluon carries colour charge; a proton (a colour singlet) carries zero net colour charge.

The Noether current table

Symmetry	Transformation	Current j^μ	Charge Q	Physical meaning
Spacetime translation	$x^\mu \rightarrow x^\mu + \epsilon^\mu$	$T^\mu{}_\nu$ (stress tensor)	$E, \mathbf{P}$	Energy, momentum
Lorentz rotation	$x \rightarrow \Lambda x$	Angular momentum tensor	$\mathbf{J}$	Angular momentum
Global $U(1)$	$\Phi \rightarrow e^{iY\alpha} \Phi$	$Y \bar{\Psi} \gamma^\mu \Psi$	Q_{em}	Electric charge
Global $SU(2)_L$	$\Psi_L \rightarrow e^{i\theta^a \tau^a/2} \Psi_L$	$\bar{\Psi}_L \gamma^\mu (\tau^a/2) \Psi_L$	T^a	Weak isospin
Global $SU(3)_c$	$q \rightarrow e^{i\theta^a T^a} q$	$\bar{q} \gamma^\mu T^a q$	C^a	Colour charge

Every row follows the same pattern: a symmetry group with continuous parameter ϵ or θ^a , a current j^μ built from the field and its conjugate contracted with the appropriate matrix (gamma matrix for the spinor index, generator matrix for the internal index), and a conserved charge that is the spatial integral of j^0 .

29.5 Noether currents beyond the free field: interactions, anomalies, and condensates

The previous subsection derived the Noether current from the Lagrangian by applying an infinitesimal symmetry transformation and reading off the conserved current from the Euler-Lagrange equations. There is a systematic habit in the textbook literature of performing this derivation using only the *free* part of the Lagrangian — the kinetic term and the mass term, without the interaction terms — and presenting the result as the correct conserved current of the full theory. This habit is understandable (the free-field current is simpler to compute and already captures the leading physics) but it is misleading in three important ways: it misses the interaction contributions to the current, it misses the quantum anomalies that modify or destroy conservation at the loop level, and it misses the condensate contributions that are essential in theories with spontaneous symmetry breaking. This subsection corrects these omissions.

The correct statement in the Fock-Born framework

In the Fock-Born framework, the primary statement of a conservation law is not the Lagrangian Noether theorem but an operator identity on the Fock space. A conserved charge $\hat{Q}$ is an operator on $\mathcal{F}$ satisfying:

$$[H, \hat{Q}] = 0, \quad (29.25)$$

where H is the *full* interacting Hamiltonian. The conservation law $[H, \hat{Q}] = 0$ is the Fock-Born statement that $\hat{Q}$ is a constant of the motion under the full Schrödinger equation: $\partial_t \hat{Q} = i[H, \hat{Q}] = 0$. The Born rendering of $\hat{Q}$ at any time t gives the same probability distribution as at time $t = 0$; the charge is protected against all dynamical processes generated by H .

The conserved current $\hat{j}^\mu(\mathbf{x}, t)$ is then defined as the local operator satisfying:

$$\hat{Q} = \int d^3\mathbf{x} \hat{j}^0(\mathbf{x}, t), \quad \partial_\mu \hat{j}^\mu(\mathbf{x}, t) = 0 \quad (\text{operator identity}), \quad (29.26)$$

where the time derivative in the continuity equation is:

$$\partial_t \hat{j}^0(\mathbf{x}, t) = i[H, \hat{j}^0(\mathbf{x}, t)], \quad (29.27)$$

computed with the full Hamiltonian H . This is the correct definition of the conserved current in the Fock-Born framework. The Noether current derived from the Lagrangian is an approximation to this operator: it is the tree-level expression, valid before quantum corrections are included. The full operator current receives corrections from interactions, from the quantum measure (anomalies), and from the ground state (condensate contributions).

Why the free-field derivation is insufficient. The Noether current from the free Lagrangian:

$$j_{\text{free}}^\mu = \frac{\partial \mathcal{L}_{\text{free}}}{\partial(\partial_\mu \phi)} \delta \phi, \quad (29.28)$$

satisfies $\partial_\mu j_{\text{free}}^\mu = 0$ when the *free* equations of motion hold. In the interacting theory, the equations of motion are modified by the interaction: $\frac{\delta S_{\text{free}}}{\delta \phi} = 0$ is replaced by $\frac{\delta S_{\text{free}}}{\delta \phi} = -\frac{\delta S_{\text{int}}}{\delta \phi}$. Applying the same Noether identity to the full Lagrangian:

$$\partial_\mu j_{\text{full}}^\mu = \partial_\mu j_{\text{free}}^\mu + \partial_\mu j_{\text{int}}^\mu = \frac{\partial \mathcal{L}_{\text{int}}}{\partial \phi} \delta \phi - \frac{\delta S_{\text{int}}}{\delta \phi} \delta \phi, \quad (29.29)$$

which vanishes when the full equations of motion hold *and* when $\mathcal{L}_{\text{int}}$ is invariant under the transformation ($\delta \mathcal{L}_{\text{int}} = 0$). If either condition fails, the free-field current is not conserved in the full theory.

Interaction contributions to the current

The electromagnetic current in QED. For the $U(1)$ transformation $\psi \rightarrow e^{i\alpha}\psi$, $A_\mu \rightarrow A_\mu - \partial_\mu\alpha/e$, the full QED Lagrangian:

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \quad (29.30)$$

is invariant under the combined transformation. Applying the Noether procedure to the full Lagrangian:

$$j_{\text{EM}}^\mu = \frac{\partial\mathcal{L}_{\text{QED}}}{\partial(\partial_\mu\psi)}\delta\psi + \frac{\partial\mathcal{L}_{\text{QED}}}{\partial(\partial_\mu A_\nu)}\delta A_\nu = \bar{\psi}\gamma^\mu\psi + \underbrace{\frac{1}{e}F^{\mu\nu}\partial_\nu\alpha}_{\text{gauge field contribution}}. \quad (29.31)$$

The second term is a total derivative $\partial_\nu(F^{\mu\nu}\alpha/e)$ when α is constant, and contributes zero to the total charge $Q = \int d^3\mathbf{x} j^0$. In this case the free-field result $j_{\text{EM}}^\mu = \bar{\psi}\gamma^\mu\psi$ is exactly correct for the charge, though the local current receives an improvement term.

The energy-momentum tensor. The energy-momentum tensor derived from the free Dirac Lagrangian is:

$$T_{\text{free}}^{\mu\nu} = \bar{\psi} i\gamma^\mu \partial^\nu \psi - \eta^{\mu\nu} \mathcal{L}_{\text{free}}. \quad (29.32)$$

This tensor has two problems: it is not symmetric ($T_{\text{free}}^{\mu\nu} \neq T_{\text{free}}^{\nu\mu}$) and it is not gauge-invariant (it contains $\partial^\nu\psi$ rather than $D^\nu\psi$). Both problems are artefacts of using the free Lagrangian. The full energy-momentum tensor from the QED Lagrangian (34.35):

$$T_{\text{QED}}^{\mu\nu} = \bar{\psi} i\gamma^\mu D^\nu \psi - F^{\mu\lambda} \partial^\nu A_\lambda - \eta^{\mu\nu} \mathcal{L}_{\text{QED}} \quad (29.33)$$

still lacks manifest symmetry. The symmetric, gauge-invariant *Belinfante-Rosenfeld tensor*:

$$\Theta^{\mu\nu} = \bar{\psi} i\gamma^{(\mu} D^{\nu)} \psi + F^{\mu\lambda} F^\nu{}_\lambda - \frac{1}{4}\eta^{\mu\nu} F_{\lambda\rho} F^{\lambda\rho} \quad (29.34)$$

is obtained by adding an improvement term $\partial_\lambda H^{\lambda\mu\nu}$ (with $H^{\lambda\mu\nu} = -H^{\mu\lambda\nu}$) to $T_{\text{QED}}^{\mu\nu}$. The Belinfante tensor is the correct physical energy-momentum tensor: it satisfies $\Theta^{\mu\nu} = \Theta^{\nu\mu}$, it is gauge-invariant, and its Born rendering gives the physical energy density and momentum density measured in experiment. It cannot be obtained from the free Lagrangian alone: the gauge field stress tensor $F^{\mu\lambda} F^\nu{}_\lambda$ is an interaction term that is absent in the free-field derivation.

In the Fock-Born framework: the Belinfante tensor is the operator $\hat{\Theta}^{\mu\nu}$ whose time component $\hat{\Theta}^{0\nu}$ generates spacetime translations on the full Fock space — it is the generator of the Poincaré group acting on $\mathcal{F}$. The condition $[H, \hat{P}^\nu] = 0$ (with $\hat{P}^\nu = \int d^3\mathbf{x} \hat{\Theta}^{0\nu}$) holds for the full QED Hamiltonian, not just for the free part.

Quantum anomalies: when symmetry survives the Lagrangian but not the measure

The most striking modification to Noether's theorem occurs at the quantum level, through the mechanism of *anomalies*. An anomaly is the failure of a classical symmetry to survive quantization: the Lagrangian is invariant under a transformation, but the quantum theory (the path integral, or equivalently the regularized operator algebra) is not.

The axial anomaly. The Dirac Lagrangian for a massless fermion is invariant under the axial (chiral) $U(1)_A$ transformation:

$$\psi \rightarrow e^{i\alpha\gamma^5}\psi, \quad \bar{\psi} \rightarrow \bar{\psi} e^{i\alpha\gamma^5}, \quad (29.35)$$

and the Noether current from the free massless Lagrangian is the axial current:

$$j_A^\mu = \bar{\psi}\gamma^\mu\gamma^5\psi, \quad (29.36)$$

which satisfies $\partial_\mu j_A^\mu = 0$ by the massless equations of motion $i\not{\partial}\psi = 0$. This suggests that the axial charge $Q_A = \int d^3\mathbf{x} j_A^0$ is conserved.

In the quantum theory, this is wrong. Fujikawa (1979) showed that the path integral measure $\mathcal{D}\psi \mathcal{D}\bar{\psi}$ is *not* invariant under the axial transformation (29.35): the Jacobian of the transformation is non-trivial and produces an additional contribution to the Ward identity. The correct quantum statement is:

$$\partial_\mu \hat{j}_A^\mu = \frac{g^2}{16\pi^2} F_{\mu\nu}^a \tilde{F}^{a\mu\nu}, \quad (29.37)$$

where $\tilde{F}^{a\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma}^a$ is the dual field strength and g is the gauge coupling. The right-hand side is the pseudoscalar density. This is the *axial anomaly* (or ABJ anomaly, after Adler, Bell, and Jackiw).

The axial charge is not conserved in the presence of gauge field configurations with non-zero topological charge:

$$\frac{dQ_A}{dt} = \int d^3\mathbf{x} \partial_\mu j_A^\mu = \frac{g^2}{16\pi^2} \int d^3\mathbf{x} F \tilde{F} = \frac{g^2}{4\pi^2} \frac{dQ_{\text{top}}}{dt}, \quad (29.38)$$

where $Q_{\text{top}} = \frac{g^2}{32\pi^2} \int d^4x F \tilde{F}$ is the topological charge (the winding number of the gauge field configuration, which changes by integers in the presence of instantons). In an instanton background of winding number $+1$, the axial charge changes by $\Delta Q_A = 2N_f$ (two units per flavour: one left-handed zero mode and one right-handed zero mode of the Dirac operator in the instanton background, by the Atiyah-Singer index theorem).

In the Fock-Born framework: the axial charge operator $\hat{Q}_A$ does not commute with the full QCD Hamiltonian H_{QCD} :

$$[H_{\text{QCD}}, \hat{Q}_A] = \frac{ig^2}{16\pi^2} \int d^3\mathbf{x} F_{\mu\nu}^a \tilde{F}^{a\mu\nu} \neq 0. \quad (29.39)$$

The free-field derivation, which gives $[H_{\text{free}}, \hat{Q}_A] = 0$, misses this commutator entirely because it misses the gauge field topology and the fermion zero modes that generate the anomaly.

Physical consequences of the axial anomaly. The anomaly (29.70) has three major physical consequences, none of which are visible from the free-field Noether current:

Pion decay $\pi^0 \rightarrow \gamma\gamma$. The neutral pion π^0 couples to the axial current $j_{A,3}^\mu = \bar{\psi}\gamma^\mu\gamma^5\tau^3\psi/2$, which is the Noether current of the broken axial isospin symmetry. The decay rate $\pi^0 \rightarrow \gamma\gamma$ is computed from the triangle diagram (a fermion loop with one axial current and two electromagnetic currents) and is entirely determined by the anomaly coefficient:

$$\Gamma(\pi^0 \rightarrow \gamma\gamma) = \frac{N_c^2 \alpha^2 m_\pi^3}{576\pi^3 f_\pi^2} \approx 7.73 \text{ eV}, \quad (29.40)$$

where $N_c = 3$ is the number of colours. The experimental value $7.82 \pm 0.14 \text{ eV}$ agrees to better than 2%. The factor N_c^2 is the anomaly coefficient squared; without the anomaly (using only the free-field current), this decay would be absent at one loop.

The η' mass. The $U(1)_A$ axial symmetry, if exact, would predict a ninth pseudo-Goldstone boson (the η') as light as the pions. The observed η' mass $m_{\eta'} \approx 958 \text{ MeV}$ is much larger than the pion mass $m_\pi \approx 140 \text{ MeV}$. The Witten-Veneziano formula:

$$m_{\eta'}^2 \approx \frac{2N_f}{f_\pi^2} \chi_{\text{top}}, \quad (29.41)$$

where $\chi_{\text{top}} = \langle Q_{\text{top}}^2 \rangle / V$ is the topological susceptibility of the QCD vacuum, shows that the η' mass is entirely due to the anomaly — specifically to the vacuum fluctuations of the topological

charge. The free-field Noether analysis predicts a massless η' ; the anomaly corrects this to 958 MeV.

The strong CP problem. If the axial $U(1)_A$ were an exact symmetry, the θ -term $\mathcal{L}_\theta = \theta g^2 F \tilde{F} / 32\pi^2$ in the QCD Lagrangian could be removed by an axial rotation of the quark fields. The anomaly (29.70) shows that this axial rotation is not a symmetry of the quantum theory: the path integral measure changes under the rotation, generating a contribution to the effective θ -angle. The strong CP problem — why $|\theta| < 10^{-10}$ is observed when there is no known symmetry reason for it to be small — is therefore a consequence of the axial anomaly.

The trace anomaly. A second important anomaly is the *trace anomaly* (or conformal anomaly). Classical massless QCD is scale-invariant: the Lagrangian $\mathcal{L}_{\text{QCD}} = -\frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu} + \bar{\psi}i\not{D}\psi$ has no dimensionful parameter for zero quark masses. The Noether current of scale invariance is the dilatation current $j_D^\mu = x_\nu T^{\mu\nu}$, and its conservation $\partial_\mu j_D^\mu = T^\mu{}_\mu = 0$ would imply a traceless energy-momentum tensor.

At the quantum level, scale invariance is broken by the renormalization procedure: the coupling $g_s(\mu)$ runs with the renormalization scale μ , introducing an implicit scale into the theory. The trace of the quantum energy-momentum tensor is:

$$\langle T^\mu{}_\mu \rangle = \frac{\beta(g_s)}{2g_s} F_{\mu\nu}^a F^{a\mu\nu} + \sum_f m_f \bar{\psi}_f \psi_f, \quad (29.42)$$

where $\beta(g_s) = -b_3 g_s^3 / (16\pi^2)$ is the QCD beta function and the sum is over all quark flavours. The first term is the trace anomaly: even for massless quarks, the quantum energy-momentum tensor has a non-zero trace proportional to the beta function. This has direct physical consequences: it is the origin of most of the proton mass. The proton mass $m_p \approx 938$ MeV receives contributions from the quark masses (only ~ 10 MeV for the up and down quarks) and from the QCD trace anomaly ($\sim 90\%$ of the proton mass), which arises from the quantum breaking of scale invariance through the running of g_s . The free-field Noether analysis, which gives $T^\mu{}_\mu = 0$ for massless fermions, misses 90% of the proton mass.

In the Fock-Born framework: the trace anomaly is the statement that $[\hat{P}^\mu, \hat{D}] \neq 0$ for the full QCD Hamiltonian, where $\hat{D} = \int d^3\mathbf{x} x_\nu \hat{\Theta}^{0\nu}$ is the dilatation operator. The free-field dilatation operator commutes with the free Hamiltonian; the anomaly is the commutator with the interaction.

Condensate contributions: the current in a broken phase

When the ground state of the theory is a condensate — a state with $\langle \Omega | \hat{\phi} | \Omega \rangle = v \neq 0$ — the Noether current derived from the free Lagrangian misses the condensate contribution to the current. This is particularly important in theories with gauge invariance, where the condensate contributes to the electromagnetic or weak current through the *diamagnetic* term.

The superconductor: paramagnetic and diamagnetic currents. In the BCS theory of superconductivity, the electromagnetic current has two contributions. The paramagnetic current is the free-field Noether current:

$$\hat{j}_{\text{para}}^\mu = -e \bar{\psi} \gamma^\mu \psi, \quad (29.43)$$

which arises from the variation of the kinetic term with respect to the gauge field. The diamagnetic current arises from the minimal coupling $(\hat{\mathbf{p}} + e\mathbf{A})^2/2m$ in the BCS Hamiltonian:

$$\hat{j}_{\text{dia}}^\mu = -\frac{e^2}{m} n_s A^\mu, \quad (29.44)$$

where $n_s = |\psi_0|^2 = N_0/V$ is the superfluid density (the condensate density). The full electromagnetic current in the superconductor is:

$$\hat{j}_{\text{SC}}^\mu = \hat{j}_{\text{para}}^\mu + \hat{j}_{\text{dia}}^\mu = -e\bar{\psi}\gamma^\mu\psi - \frac{e^2 n_s}{m} A^\mu. \quad (29.45)$$

The diamagnetic term $-e^2 n_s A^\mu/m$ is the *London equation*: it is the origin of the Meissner effect (the expulsion of magnetic flux from a superconductor) and the zero electrical resistance. Applying the London equation to the Maxwell equations:

$$\nabla^2 \mathbf{B} = \frac{e^2 n_s}{m} \mathbf{B} = \frac{\mathbf{B}}{\lambda_L^2}, \quad \lambda_L = \sqrt{\frac{m}{e^2 n_s}}, \quad (29.46)$$

the magnetic field decays exponentially inside the superconductor with the London penetration depth λ_L . The diamagnetic current is entirely a condensate effect: it is proportional to the superfluid density $n_s = N_0/V$, which vanishes above T_c (no condensate, $N_0 = 0$) and equals the total electron density at $T = 0$ in a perfect BCS superconductor. The free-field Noether current $\hat{j}_{\text{para}}^\mu$ alone gives no Meissner effect, no London equation, and no superconductivity.

In the Fock-Born framework: the diamagnetic current arises from the Born rendering of the current operator $\hat{j}^\mu = \delta H / \delta A_\mu$ in the BCS ground state $|\Omega_{\text{BCS}}\rangle$. The BCS ground state is a condensate state in which the zero-momentum Cooper pair mode is macroscopically occupied; its contribution to the current expectation value is the diamagnetic term. The free-field derivation uses the vacuum $|0\rangle$ rather than the BCS ground state $|\Omega_{\text{BCS}}\rangle$, and $\langle 0 | \hat{j}_{\text{dia}}^\mu | 0 \rangle = 0$ because the vacuum has no condensate.

The Higgs mechanism: the W and Z currents. In the electroweak theory, the same structure appears for the W and Z bosons. The free-field $SU(2)_L$ Noether current is:

$$j_{\text{free}}^{\mu,i} = \bar{\psi}_L \gamma^\mu \tau^i \psi_L, \quad (29.47)$$

where τ^i are the $SU(2)_L$ generators. After the Higgs condensate develops ($\langle \phi \rangle = v/\sqrt{2}$), the Higgs kinetic term $|D_\mu \phi|^2$ contributes a diamagnetic current:

$$j_{\text{Higgs}}^{\mu,i} = -gv^2 W^{\mu,i} + \dots \quad (29.48)$$

which gives the W boson a mass $M_W = gv/2$. This is the Hamiltonian-level statement of the Higgs mechanism: the condensate contributes a diamagnetic current to the $SU(2)_L$ Noether current, which appears as a mass term for the W boson in the same way that the diamagnetic current appears as a mass term for the photon in a superconductor (the Meissner mass). The Higgs mechanism is the relativistic Meissner effect.

The free-field Noether current (29.47) misses the condensate contribution (29.48) entirely. Deriving the W mass from the free-field current alone is impossible: the mass is a condensate effect, visible only when the Born rendering is applied to the interacting ground state $|\Omega_{\text{Higgs}}\rangle$ rather than the free vacuum $|0\rangle$.

Summary: three levels of the Noether current

Level	Current	What it captures	What it misses
Free field	$j_{\text{free}}^\mu = \frac{\partial \mathcal{L}_{\text{free}}}{\partial(\partial_\mu \phi)} \delta\phi$	Tree-level charges; leading-order Ward identities	Interaction corrections; anomalies; condensate contributions
Full classical	$j_{\text{full}}^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta\phi$ (from full $\mathcal{L}$)	Interaction corrections; Belinfante tensor; London equation	Quantum anomalies (axial, trace); higher-loop Ward identities
Full quantum	$\hat{j}^\mu$ with $[H, \hat{Q}] = 0$ for full interacting H	Everything: interactions, anomalies, condensate, all loops	Nothing: this is the correct current

The standard textbook Noether current sits at the first row of this table. The physical current — the one whose Born rendering gives the measured charge distribution, the measured electromagnetic response, and the measured decay rates — sits at the third row. The distance between the first and the third row is the distance between the free-field approximation and the full quantum theory, and it contains some of the most important physics in the Standard Model: 90% of the proton mass (trace anomaly), the decay rate of the π^0 (axial anomaly), the strong CP problem (anomalous $U(1)_A$), the Meissner effect (diamagnetic current), the W and Z masses (Higgs condensate current), and the Josephson effect (condensate phase current).

The Fock-Born framework makes the correct statement unambiguous: the conserved charge is the operator $\hat{Q}$ satisfying $[H, \hat{Q}] = 0$ for the full Hamiltonian H . The free-field Noether theorem is the first approximation to this operator, valid in the non-interacting limit. Every correction beyond this limit — from interactions, from the quantum measure, from the ground state — is a physical effect that modifies the current and can be measured in experiment. Teaching the Noether theorem from the free Lagrangian alone, and presenting the result as the conserved current of the full theory, is teaching the first approximation as the exact answer. It is a disservice comparable to teaching Newtonian gravity and presenting it as the exact theory of gravitational waves.

From global to local: what Noether does not give

Noether’s theorem applies to *global* symmetries: the parameter α or θ^a is constant across spacetime. It gives a conserved current and a conserved charge.

When the symmetry is *local* ($\alpha = \alpha(x)$, varying from point to point), Noether’s theorem as stated above does not directly apply, because a local symmetry is not really a symmetry in the usual sense—it is a *redundancy* of description. The field configurations related by a local gauge transformation are physically identical; they are the same state described in different “gauges” (different local phase conventions).

The consequence of a local gauge symmetry is not a new conserved charge (the global version already gave that) but rather the existence of a *gauge boson*: the compensating vector field A_μ that we constructed in the previous section. The local symmetry enforces a *constraint* (Gauss’s law in electromagnetism, or its non-Abelian generalisations), and this constraint is what prevents the longitudinal polarisation of the gauge boson from propagating as a physical degree of freedom, keeping the photon with two transverse polarisations and the gluons likewise.

The logical chain is therefore:

$$\underbrace{\text{global symmetry}}_{\alpha \text{ constant}} \xrightarrow{\text{Noether}} \underbrace{\text{conserved current and charge}}_{\partial_\mu j^\mu = 0, Q = \text{const}} \xrightarrow{\text{promote to local } \alpha \rightarrow \alpha(x)} \xrightarrow{\text{gauge principle}} \underbrace{\text{gauge boson and force } A_\mu}$$

Global symmetry gives conservation laws. Local symmetry gives forces. Together they are the complete dictionary between the abstract mathematics of group theory and the observable content of particle physics.

29.6 The Wald identity and masslessness protection

The previous subsection stated that a local gauge symmetry implies a constraint. We now make this precise: gauge invariance implies an *algebraic identity among the equations of motion*—a differential relation that holds identically, not just on shell. This identity is the deepest consequence of gauge symmetry, and it is what protects the masslessness of gauge bosons against all dynamical corrections.

The identity. Consider an action $S[\phi^i]$ (where ϕ^i denotes all fields collectively) that is invariant under a gauge transformation $\phi^i \rightarrow \phi^i + \delta_\alpha \phi^i$ for arbitrary spacetime-dependent parameter $\alpha(x)$:

$$\delta_\alpha S = \int d^4x \frac{\delta S}{\delta \phi^i(x)} \delta_\alpha \phi^i(x) = 0 \quad \text{for all } \alpha(x). \quad (29.49)$$

The Euler–Lagrange equations of motion are $E_i(x) \equiv \delta S / \delta \phi^i(x) = 0$. The gauge variation of each field is a differential operator acting on α : $\delta_\alpha \phi^i(x) = \mathcal{R}^i[\alpha](x)$, where $\mathcal{R}^i$ is linear in α and its derivatives. Substituting into (29.49):

$$\int d^4x E_i(x) \mathcal{R}^i[\alpha](x) = 0 \quad \text{for all } \alpha(x). \quad (29.50)$$

Since $\alpha(x)$ is *arbitrary*, we can integrate by parts to move all derivatives off α and onto the equations of motion E_i . The result is:

$$\boxed{\mathcal{R}^{i\dagger} E_i(x) \equiv 0}, \quad (29.51)$$

where $\mathcal{R}^{i\dagger}$ is the formal adjoint of $\mathcal{R}^i$. This is an *identity*: it holds for all field configurations, not just solutions of $E_i = 0$. It is a differential relation among the equations of motion, forced by gauge invariance.

Electromagnetism. The $U(1)$ gauge transformation of the photon field: $\delta_\alpha A_\mu = \partial_\mu \alpha$. The equation of motion: $E^\mu = \partial_\nu F^{\nu\mu} - j^\mu$. The identity (29.51) gives:

$$\partial_\mu E^\mu = \partial_\mu \partial_\nu F^{\nu\mu} - \partial_\mu j^\mu \equiv 0. \quad (29.52)$$

The first term vanishes identically (antisymmetry of $F^{\nu\mu} = -F^{\mu\nu}$): $\partial_\mu \partial_\nu F^{\nu\mu} = 0$. Therefore $\partial_\mu j^\mu = 0$: current conservation is an algebraic consequence of gauge invariance, not an additional assumption. This is the same conservation law that Noether’s theorem gives for the global $U(1)$ symmetry, but here it follows from the *local* structure automatically.

General relativity. The diffeomorphism gauge transformation of the metric: $\delta_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu$. The equation of motion: $E^{\mu\nu} = G^{\mu\nu} - 8\pi G T^{\mu\nu}/c^4$. The identity (29.51) gives:

$$\nabla_\mu E^{\mu\nu} \equiv 0 \quad \implies \quad \nabla_\mu G^{\mu\nu} \equiv 0, \quad (29.53)$$

which is the contracted Bianchi identity — here derived not from differential geometry but from the gauge invariance of the action. Combined with the Einstein equation: $\nabla_\mu T^{\mu\nu} = 0$ (stress-energy conservation). The geodesic equation for matter follows. Wheeler’s “two instructions” are one identity.

Masslessness protection. The Wald identity has a direct physical consequence: it protects the masslessness of the gauge boson.

Consider adding a mass term $\frac{1}{2}m^2 A_\mu A^\mu$ to the Maxwell Lagrangian. Under the gauge transformation $\delta_\alpha A_\mu = \partial_\mu \alpha$:

$$\delta_\alpha \left(\frac{1}{2} m^2 A_\mu A^\mu \right) = m^2 A^\mu \partial_\mu \alpha \neq 0. \quad (29.54)$$

The mass term breaks gauge invariance. The Wald identity (29.51) is violated. The consequence is immediate and severe:

The constraint is lost. Without gauge invariance, the identity $\partial_\mu E^\mu \equiv 0$ no longer holds. The Gauss law $\nabla \cdot \mathbf{E} = \rho$ is no longer a constraint (an identity among equations of motion) but an independent equation of motion. The field A_0 acquires its own dynamics. The number of propagating degrees of freedom increases from 2 (transverse polarisations) to 3 (two transverse + one longitudinal).

The longitudinal mode is dangerous. The massive vector propagator:

$$\tilde{\Delta}^{\mu\nu}(p) = \frac{-i}{p^2 - m^2} \left(\eta^{\mu\nu} - \frac{p^\mu p^\nu}{m^2} \right). \quad (29.55)$$

The $p^\mu p^\nu / m^2$ term is the longitudinal polarisation. In the limit $m \rightarrow 0$: this term diverges. Scattering amplitudes involving longitudinal polarisations grow as E^2/m^2 at high energies, violating unitarity at $E \sim m/g$ (where g is the gauge coupling). A massive photon with $m = 0.1 \text{ eV}$ and $g = e \approx 0.3$ would violate unitarity at $E \sim 0.3 \text{ eV}$ — far below the energies where QED is tested.

The Wald identity prevents this: gauge invariance is exact $\rightarrow$ the identity holds $\rightarrow$ the constraint eliminates the longitudinal mode $\rightarrow$ the gauge boson has exactly 2 polarisations $\rightarrow$ the mass is exactly zero. The chain is algebraic: it depends on the symmetry of the action, not on the dynamics (the specific interactions, the coupling constants, the loop order).

The kinetic–dynamic independence. This is the deepest lesson of the Wald identity: the *kinetic structure* of a gauge field — its propagator, its number of physical polarisations, its mass (zero or nonzero) — is determined entirely by the gauge symmetry of the action, independent of the *dynamical details* (the specific interactions, the matter content, the values of the coupling constants, the non-perturbative effects).

The reason is that the Wald identity (29.51) is an *off-shell* identity: it holds for all field configurations, not just solutions. It constrains the structure of the equations of motion algebraically, before any solution is found. The propagator (the inverse of the kinetic operator) inherits this constraint: its pole structure (the mass) and its tensor structure (the polarisations) are fixed by the identity.

At the quantum level, the Wald identity becomes the *Ward–Takahashi identity*:

$$q_\mu \Gamma^{\mu\alpha\beta}(q, p_1, p_2) = e [S_F^{-1}(p_1) - S_F^{-1}(p_2)]^{\alpha\beta}, \quad (29.56)$$

where $\Gamma^{\mu\alpha\beta}$ is the full (all-loop) vertex function and S_F^{-1} is the full (all-loop) inverse fermion propagator. This identity holds at every order in perturbation theory and non-perturbatively. Its consequence for the photon propagator: the self-energy is transverse,

$$\Pi^{\mu\nu}(q) = (q^2 \eta^{\mu\nu} - q^\mu q^\nu) \Pi(q^2), \quad (29.57)$$

with no $\eta^{\mu\nu}$ piece that could generate a mass. The full photon propagator:

$$\tilde{\Delta}_{\text{full}}^{\mu\nu}(q) = \frac{-i\eta^{\mu\nu}}{q^2[1 - \Pi(q^2)] + i\varepsilon} + \text{gauge-dependent terms}. \quad (29.58)$$

The pole remains at $q^2 = 0$ regardless of $\Pi(q^2)$: the photon is massless to all orders, because the transversality (29.57) forbids a mass-generating correction.

This is kinetic–dynamic independence: the kinetic structure (mass = 0, two polarisations, transverse propagator) is locked by the Ward–Takahashi identity, while the dynamics (the running coupling $\alpha(\mu)$, the vacuum polarisation function $\Pi(q^2)$, the Lamb shift, the anomalous magnetic moment) are free to vary. The dynamics determine *how strongly* the photon couples at each energy scale. The kinematics determine *what the photon is*: a massless, transverse, two-polarisation excitation. The two are decoupled.

Non-Abelian generalisation. For non-Abelian gauge theories ($SU(3)_c$, $SU(2)_L$), the Wald identity generalises to the Slavnov–Taylor identities:

$$q^\mu \tilde{\Delta}_{\mu\nu}^{ab}(q) = (\text{ghost propagator terms}), \quad (29.59)$$

which are more complex (involving ghost Green’s functions) but serve the same purpose: they constrain the gluon self-energy to be transverse and the gluon mass to be zero, at every order in perturbation theory. This is why the gluon is massless despite the strong coupling at low energies: confinement changes the dynamics (no free gluons at long distances) but does not change the kinematics (the gluon mass is protected by the Slavnov–Taylor identity at every scale).

The Higgs mechanism: controlled violation. The Higgs mechanism (section ??) is the only known way to give a gauge boson a mass without destroying the Wald identity. The trick: the gauge symmetry is spontaneously broken, not explicitly broken. The Lagrangian remains gauge-invariant (the Wald identity holds), but the vacuum state $|\Omega\rangle$ does not respect the symmetry. The Goldstone bosons $\theta^a(x)$ of the broken generators are “eaten” by the gauge bosons, becoming their longitudinal polarisations. The result: the gauge boson acquires a mass $M_W = gv/2$, the longitudinal mode is a physical degree of freedom (the eaten Goldstone), and the theory remains unitary because the Goldstone’s kinetic term provides the correct high-energy behaviour.

In the Fock-Born language: the Wald identity (a property of the Fock-layer action) forbids an explicit mass. The Higgs condensate (a property of the Fock-layer vacuum state) provides a mass through spontaneous breaking. The kinetic structure is modified (from 2 to 3 polarisations) in a way that is compatible with the identity, because the additional polarisation comes from a physical field (the Goldstone boson), not from a forbidden longitudinal mode.

The Fock-Born reading. The Wald identity is the strongest example of the engine’s architecture determining a Born-layer output exactly:

Engine property	Born-layer consequence	Protection mechanism
$U(1)_{\text{em}}$ gauge inv. of S	$m_\gamma = 0$ exactly	Wald $\rightarrow$ Ward–Takahashi $\rightarrow \Pi^{\mu\nu}$ transverse
$SU(3)_c$ gauge inv. of S	$m_g = 0$ exactly	Wald $\rightarrow$ Slavnov–Taylor
$SU(2)_L$ gauge inv. of S	m_W, m_Z from Higgs VEV only	Wald + spontaneous breaking
Diffeo. inv. of S_{EH}	$m_{\text{graviton}} = 0$ exactly	Wald $\rightarrow$ Bianchi identity

The masses in the left column are Born-layer observables: what a spectrometer measures. Their values are determined by the Fock-layer action’s symmetry, not by its dynamics. No amount of strong coupling, non-perturbative effects, or loop corrections can change $m_\gamma = 0$ or $m_g = 0$, because the protection is algebraic (the Wald identity), not dynamical (a cancellation between diagrams).

Compare with the pNGB mass (section ??): m_π is protected by an *approximate* symmetry of the engine, so it is small but nonzero ($m_\pi^2 \propto m_q$). The gauge boson mass is protected by an

exact symmetry of the engine, so it is exactly zero. Approximate symmetry gives approximate protection (light pNGBs); exact symmetry gives exact protection (massless gauge bosons). The hierarchy of masses in the Standard Model is a hierarchy of symmetry protections in the Fock-layer engine.

29.7 Discrete symmetries: parity, pseudoscalars, and pseudovectors

The representation theory developed in the previous subsections classified fields by their behaviour under *proper orthochronous* Lorentz transformations: rotations and boosts, continuously connected to the identity. The full Lorentz group $O(1,3)$ also contains *discrete* transformations that are not continuously connected to the identity: *parity* P (spatial reflection $\mathbf{x} \rightarrow -\mathbf{x}$, $t \rightarrow t$), *time reversal* T ($\mathbf{x} \rightarrow \mathbf{x}$, $t \rightarrow -t$), and their combination PT . Including these discrete transformations refines the classification: representations that were indistinguishable under the proper Lorentz group split into pairs distinguished by their behaviour under parity. This doubling produces the scalar/pseudoscalar pair, the vector/pseudovector pair, and ultimately the $V - A$ structure of the weak interaction.

Parity as a Lorentz transformation

Parity is the transformation:

$$P : (t, \mathbf{x}) \mapsto (t, -\mathbf{x}), \quad (29.60)$$

which reverses all three spatial coordinates while leaving time unchanged. In four-vector notation, P acts as:

$$P^\mu{}_\nu = \text{diag}(+1, -1, -1, -1), \quad (29.61)$$

the same matrix as the Minkowski metric $\eta_{\mu\nu}$ (up to sign conventions). Parity is not a proper Lorentz transformation because $\det P = -1$, not $+1$; it cannot be reached by continuously deforming the identity. Under proper Lorentz transformations alone, fields are classified by their spin (j_L, j_R) ; including parity, each representation (j_L, j_R) with $j_L \neq j_R$ pairs with (j_R, j_L) (parity swaps left and right), and representations with $j_L = j_R$ split into two parity eigenspaces labelled by $P = \pm 1$.

Scalars and pseudoscalars

A field with no Lorentz indices and spin $(0,0)$ can transform in two ways under parity.

A *scalar* field $\phi(x)$ transforms as:

$$\phi(t, \mathbf{x}) \xrightarrow{P} +\phi(t, -\mathbf{x}). \quad (29.62)$$

The field value does not change; only the evaluation point is reflected. This is the $(0,0)^+$ representation.

A *pseudoscalar* field $\varphi(x)$ transforms as:

$$\varphi(t, \mathbf{x}) \xrightarrow{P} -\varphi(t, -\mathbf{x}). \quad (29.63)$$

The field value picks up an additional minus sign. This is the $(0,0)^-$ representation. Under proper Lorentz transformations, scalars and pseudoscalars are completely indistinguishable: both transform with $S(\Lambda) = 1$. Parity is the only transformation that separates them.

Physical examples of pseudoscalars. The pion π^a ($a \in \{1, 2, 3\}$ for π^+, π^-, π^0) is the most important pseudoscalar in the Standard Model. It is a composite field — a bound state of a quark and an antiquark — whose pseudoscalar nature follows from the quark bilinear structure

$\bar{\psi}\gamma^5\psi$: the γ^5 inserts the minus sign under parity. The pion is the (approximate) Goldstone boson of spontaneous chiral symmetry breaking in QCD: the chiral condensate $\langle\bar{\psi}\psi\rangle \neq 0$ breaks $SU(2)_L \times SU(2)_R$ to $SU(2)_V$, and the three pions are the three pseudo-Goldstone bosons corresponding to the three broken axial generators. They are pseudoscalars because the broken generators are themselves odd under parity.

The QCD axion $a(x)$ is an elementary pseudoscalar proposed by Peccei and Quinn to solve the strong CP problem. Its coupling to QCD:

$$\mathcal{L}_{\text{axion}} \supset \frac{a}{f_a} \frac{g_s^2}{32\pi^2} F_{\mu\nu}^a \tilde{F}^{a\mu\nu}, \quad (29.64)$$

where $\tilde{F}^{a\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma}^a$ is the dual field strength, requires $a(x)$ to be a pseudoscalar: the combination $F\tilde{F}$ is a pseudoscalar (because $\epsilon^{\mu\nu\rho\sigma}$ is a pseudotensor), and for (29.64) to be a Lorentz scalar in the Lagrangian, $a(x)$ must also be a pseudoscalar to cancel the minus sign.

Vectors and pseudovectors (axial vectors)

A field carrying one Lorentz index μ and transforming in the $(\frac{1}{2}, \frac{1}{2})$ representation can also split into two parity eigenspaces.

A *vector* field $V^\mu(x)$ (polar vector) transforms as:

$$V^\mu(t, \mathbf{x}) \xrightarrow{P} P^\mu{}_\nu V^\nu(t, -\mathbf{x}) = (V^0(t, -\mathbf{x}), -\mathbf{V}(t, -\mathbf{x})). \quad (29.65)$$

The spatial components reverse sign; the time component does not. The electric field $\mathbf{E}$, the momentum $\mathbf{p}$, and the electromagnetic four-potential A^μ (the photon field) are polar vectors.

A *pseudovector* (axial vector) field $A^\mu(x)$ transforms as:

$$A^\mu(t, \mathbf{x}) \xrightarrow{P} -P^\mu{}_\nu A^\nu(t, -\mathbf{x}) = (-A^0(t, -\mathbf{x}), +\mathbf{A}(t, -\mathbf{x})). \quad (29.66)$$

The spatial components do not reverse; the time component does. This is the $(\frac{1}{2}, \frac{1}{2})^-$ representation. The magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$ is a pseudovector: the curl introduces a Levi-Civita symbol ϵ_{ijk} which is itself a pseudotensor, giving the extra minus sign. The angular momentum $\mathbf{L} = \mathbf{x} \times \mathbf{p}$ is a pseudovector for the same reason.

Physical intuition. If you stand in front of a mirror, your right hand becomes a left hand (parity reversal): your velocity (a polar vector) reverses, but your angular momentum (a pseudovector) does not, because both $\mathbf{x}$ and $\mathbf{p}$ reverse under parity, and their cross product $\mathbf{x} \times \mathbf{p}$ does not. Similarly, a current loop reverses direction under parity (the current is a polar vector), but the magnetic field it produces does not reverse (it is a pseudovector).

The Levi-Civita tensor as a pseudotensor

The origin of all pseudo-representations is the *Levi-Civita pseudotensor* $\epsilon^{\mu\nu\rho\sigma}$, which is totally antisymmetric and transforms as:

$$\epsilon^{\mu\nu\rho\sigma} \xrightarrow{P} -\epsilon^{\mu\nu\rho\sigma}. \quad (29.67)$$

It picks up a minus sign under parity (and under any improper Lorentz transformation with $\det \Lambda = -1$) because its definition involves the sign of a permutation, and reversing all three spatial coordinates is an odd permutation. Any object contracted with an odd number of Levi-Civita tensors is a pseudo-object: a scalar contracted with one ϵ becomes a pseudoscalar; a vector contracted with one ϵ becomes a pseudovector. This is the algebraic origin of all pseudo-representations.

The complete table of Dirac bilinears

The most important application of this classification in the Standard Model is the decomposition of all possible Lorentz-covariant bilinears that can be built from a Dirac spinor Ψ . Since Ψ has four complex components, the space of 4×4 matrices acting between $\bar{\Psi}$ and Ψ is 16-dimensional. This 16-dimensional space decomposes into five irreducible Lorentz representations, each with a definite parity:

Bilinear	Name	Lorentz rep	Parity	Dim	SM role
$\bar{\Psi}\Psi$	Scalar	$(0,0)^+$	+1	1	Dirac mass, Yukawa
$\bar{\Psi}\gamma^5\Psi$	Pseudoscalar	$(0,0)^-$	-1	1	Axion coupling, $\theta F\tilde{F}$
$\bar{\Psi}\gamma^\mu\Psi$	Vector	$(\frac{1}{2}, \frac{1}{2})^+$	+1	4	EM current, Z coupling
$\bar{\Psi}\gamma^\mu\gamma^5\Psi$	Axial vector	$(\frac{1}{2}, \frac{1}{2})^-$	-1	4	Weak $V - A$ current
$\bar{\Psi}\sigma^{\mu\nu}\Psi$	Tensor	$(1,0) \oplus (0,1)$	mixed	6	Magnetic moment

The total dimension is $1 + 1 + 4 + 4 + 6 = 16$, spanning the complete space of 4×4 matrices. This is the Fierz completeness relation: every Lorentz-covariant combination of two Dirac spinors can be expressed as a linear combination of these five structures. In the Fock-Born framework, each row of this table is a different choice of observable $\hat{O} = \bar{\Psi}\Gamma\Psi$ (where $\Gamma \in \{1, \gamma^5, \gamma^\mu, \gamma^\mu\gamma^5, \sigma^{\mu\nu}\}$), and the Born rendering of each gives a different classical quantity with a different transformation property under parity.

The $V - A$ structure of the weak interaction

The most consequential application of this classification in the Standard Model is the structure of the weak charged current. The $W^\pm$ boson couples to the current:

$$J_W^\mu = \bar{\Psi} \gamma^\mu \frac{1 - \gamma^5}{2} \Psi = \frac{1}{2}(\bar{\Psi}\gamma^\mu\Psi - \bar{\Psi}\gamma^\mu\gamma^5\Psi) = \frac{1}{2}(J_V^\mu - J_A^\mu), \quad (29.68)$$

a combination of vector and axial vector currents with equal coefficients. This is the $V - A$ (vector minus axial vector) structure. The projector $(1 - \gamma^5)/2$ is the left-handed projector: it projects the Dirac spinor onto its left-handed Weyl component $(\psi_L)_s$. The W boson therefore couples only to left-handed fermions and right-handed antifermions, not to right-handed fermions or left-handed antifermions.

Under parity, $J_V^\mu \rightarrow +J_V^\mu$ and $J_A^\mu \rightarrow -J_A^\mu$, so:

$$J_W^\mu = \frac{1}{2}(J_V^\mu - J_A^\mu) \xrightarrow{P} \frac{1}{2}(J_V^\mu + J_A^\mu) = J_R^\mu, \quad (29.69)$$

the right-handed current. The weak interaction transforms under parity into a theory that couples only to right-handed fermions: a theory that does not exist in nature. This is maximal parity violation: the weak interaction violates parity as strongly as possible.

In the Fock-Born framework: the Fock basis vectors of the weak interaction are labelled by chirality (L or R) as well as momentum and spin. The W boson creation/annihilation operators couple only to the L -labelled Fock states. The Born rendering of the chirality observable separates

left-handed from right-handed states, and the weak interaction is invisible to the right-handed sector. Parity would swap the L and R labels on all Fock basis vectors, transforming the Standard Model into a parity-mirror theory that is physically different. The Standard Model is not parity invariant because the L and R sectors of the Fock space are not treated symmetrically by the Hamiltonian.

The axial anomaly and the strong CP problem

The axial vector current $J_A^\mu = \bar{\Psi}\gamma^\mu\gamma^5\Psi$ appears to be conserved classically in a theory of massless fermions: $\partial_\mu J_A^\mu = 0$ by the equations of motion. But at the quantum level, this conservation is violated by an *anomaly*: the path integral measure $\mathcal{D}\Psi\mathcal{D}\bar{\Psi}$ is not invariant under axial rotations $\Psi \rightarrow e^{i\alpha\gamma^5}\Psi$, and the violation produces a non-zero divergence:

$$\partial_\mu J_A^\mu = \frac{g_s^2}{16\pi^2} F_{\mu\nu}^a \tilde{F}^{a\mu\nu}, \quad (29.70)$$

where the right-hand side is the pseudoscalar combination $F\tilde{F}$. This is the *axial anomaly* (or chiral anomaly), one of the most important quantum effects in the Standard Model. It has three major consequences.

First, it explains why the η' meson is heavy: the $U(1)_A$ axial symmetry that would otherwise predict a ninth light Goldstone boson is broken by the anomaly (29.70), giving the η' a mass of 958 MeV, much larger than the other pseudo-Goldstone bosons.

Second, it generates the *strong CP problem*: the QCD Lagrangian can contain a term:

$$\mathcal{L}_\theta = \frac{\theta g_s^2}{32\pi^2} F_{\mu\nu}^a \tilde{F}^{a\mu\nu}, \quad (29.71)$$

which is a Lorentz scalar (all indices contracted) but a pseudoscalar under parity (it changes sign under P). A non-zero θ would generate a neutron electric dipole moment of order $d_n \sim 10^{-16} \theta e \cdot \text{cm}$, but experiments constrain $|d_n| < 1.8 \times 10^{-26} e \cdot \text{cm}$, giving $|\theta| \lesssim 10^{-10}$: the strong CP problem is the question of why θ is so small. The Peccei-Quinn axion promotes θ to a dynamical pseudoscalar field $a(x)/f_a$ that relaxes to zero dynamically, solving the problem.

Third, the anomaly is responsible for the decay $\pi^0 \rightarrow \gamma\gamma$: the neutral pion couples to two photons through the axial anomaly, with a rate that can be computed exactly from (29.70) and agrees with experiment to 1% precision.

Time reversal and charge conjugation

For completeness, the other two discrete symmetries of the Lorentz group deserve mention.

Time reversal $T : t \rightarrow -t, \mathbf{x} \rightarrow \mathbf{x}$ is an antiunitary transformation (it involves complex conjugation of the wavefunction as well as the coordinate reversal), which makes it more subtle than parity. Under T , the electric field is even ($\mathbf{E} \rightarrow +\mathbf{E}$, because T reverses both the charge motion and the retardation) and the magnetic field is odd ($\mathbf{B} \rightarrow -\mathbf{B}$, because T reverses the current that produces it).

Charge conjugation C replaces each particle with its antiparticle: $\psi \rightarrow \psi^c = C\bar{\psi}^T$, where C is the charge conjugation matrix satisfying $C\gamma^\mu C^{-1} = -(\gamma^\mu)^T$. Charge conjugation reverses the sign of all additive quantum numbers (electric charge, baryon number, lepton number) without changing spacetime.

The *CPT theorem* states that any local, Lorentz-invariant, unitary quantum field theory is invariant under the combined transformation CPT . This is not a postulate but a theorem: it follows from the axioms of quantum field theory (locality, Lorentz invariance, unitarity) and

cannot be violated without breaking one of these axioms. The Standard Model violates C , P , and CP individually (the weak interaction violates both C and P maximally, and the CKM matrix introduces a small CP -violating phase), but is invariant under CPT as required.

Summary: the complete classification. Including the discrete symmetries P , T , and C alongside the proper Lorentz representations gives the complete classification of all field types in the Standard Model:

Field type	Lorentz rep	P	C	T	SM example
Scalar	$(0, 0)^+$	+1	+1	+1	Higgs ϕ
Pseudoscalar	$(0, 0)^-$	-1	+1	-1	Pion π , axion a
Vector	$(\frac{1}{2}, \frac{1}{2})^+$	+1	-1	-1	Photon A^μ , Z^0
Axial vector	$(\frac{1}{2}, \frac{1}{2})^-$	-1	-1	+1	Axial current $\bar{\Psi}\gamma^\mu\gamma^5\Psi$
Left Weyl spinor	$(\frac{1}{2}, 0)$	swaps to R	$\psi_L \rightarrow \psi_R^c$	$\psi_L \rightarrow \psi_L^*$	Left-handed quarks, leptons
Right Weyl spinor	$(0, \frac{1}{2})$	swaps to L	$\psi_R \rightarrow \psi_L^c$	$\psi_R \rightarrow \psi_R^*$	Right-handed quarks, leptons
Dirac spinor	$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	$\Psi \rightarrow \gamma^0\Psi$	$\Psi \rightarrow C\bar{\Psi}^T$	antiunitary	Electron, quarks

The Fock basis vectors of the Standard Model carry all of these labels: momentum $\mathbf{k}$, spin s or $\dot{s}$, colour α , isospin i , species f , particle/antiparticle $\pm$, and the parity and charge conjugation quantum numbers implicit in the representation assignment. The Born rendering layer selects which of these labels to measure; the Hamiltonian preserves the CPT combination of all of them. Every term in the Standard Model Lagrangian is built from the bilinear structures of the Dirac table above, contracted with the appropriate invariant tensors, and the discrete symmetry properties of each term determine which interactions preserve C , P , T , and CP individually.

30 The Dance between Matter and Force: Local Gauge Invariance

30.1 Force fields as messengers for dimension reduction

The Noether current analysis of the previous section showed that every continuous global symmetry produces a conserved current. But global transformations — the same matrix U applied everywhere in spacetime simultaneously — are an idealisation. A more honest question is: what happens when the transformation is allowed to vary from point to point? This is the question of local gauge invariance, and the answer leads directly to the existence of all forces in the Standard Model.

The problem: the derivative sees the difference. A matter field $\psi^i(x)$ carrying an internal index i can be rotated at each spacetime point by a different $SU(N)$ matrix $U(x)$. The field value at x rotates by $U(x)$; the field value at a nearby point $x + dx$ rotates by a slightly different matrix $U(x + dx)$. The derivative $\partial_\mu \psi^i$ compares the field at two neighbouring points, and because the two points are rotated by different matrices, the comparison is corrupted: the derivative picks up an extra term proportional to $\partial_\mu U(x)$ that does not transform covariantly. The Lagrangian, built from $\partial_\mu \psi$, is therefore not invariant under local gauge transformations, even if it was perfectly invariant under global ones.

The solution: a force field as a messenger. The remedy is to introduce a *gauge field* $A_\mu^a(x)$ that acts as a messenger between neighbouring spacetime points. The gauge field is specifically designed to transform under local $SU(N)$ transformations in a way that exactly cancels the extra term from $\partial_\mu U(x)$. Replacing the ordinary derivative ∂_μ with the covariant derivative $D_\mu = \partial_\mu - igA_\mu^a T^a$ restores the covariant transformation of the derivative term, and the Lagrangian is once again locally gauge invariant. The force field was not introduced by choice; the matter field’s kinetic term demanded it. Every matter field carrying a local gauge index requires a companion force field to make its derivative well-defined across spacetime. Matter demands the force.

Multiple indices: multiple force fields. When the field carries multiple internal indices — as Standard Model fields do, with colour index α for $SU(3)_c$, isospin index i for $SU(2)_L$, and hypercharge for $U(1)_Y$ — each index transforms under its own independent local gauge group, and each requires its own separate force field to compensate. The $SU(3)_c$ colour index demands the eight gluons G_μ^a ; the $SU(2)_L$ isospin index demands the three W bosons W_μ^i ; the $U(1)_Y$ hypercharge demands the single B boson B_μ . The three forces of the Standard Model — strong, weak, and electromagnetic — arise from three independent local gauge invariances, each with its own messenger field, each compensating for the variation of a different internal index across spacetime.

Gauge invariance is not a symmetry: it is a redundancy. Before going further, it is essential to resist the most common narrative in textbooks: that local gauge invariance is a physical symmetry of nature, and that this symmetry “generates” or “demands” the force fields. This framing is misleading in a fundamental way.

A true physical symmetry relates different physical situations that happen to be equivalent: Lorentz invariance says that physics looks the same in all inertial frames, and this is a non-trivial, testable, empirical fact about the world. Local gauge invariance is something entirely different. Two field configurations $\psi^i(x), A_\mu^a(x)$ and $\psi'^i(x), A'_\mu^a(x)$ related by a local gauge transformation are not two different physical situations that happen to give the same predictions: they are *the same physical situation*, described in two different mathematical languages. The gauge transformation is a redundancy in our description, not a symmetry of nature.

The correct framing is therefore not “symmetry demands the force” but “redundancy enables dimension reduction.” The gauge orbit $[A_\mu]$ — the set of all field configurations related to A_μ by local gauge transformations — represents a single physical state. The physical configuration space is not the space of all fields but the space of gauge orbits: a dramatically smaller space. The force field A_μ^a is introduced precisely to make this orbit structure well-defined. Without it, the derivative of the matter field is not gauge-covariant, the orbit structure is ill-defined, and the dimension reduction cannot be performed.

Dimension reduction and good behaviour: asymptotic freedom. The dimension reduction from fields to gauge orbits is not merely a formal nicety. It has profound physical consequences. The larger the gauge orbit — the more field configurations that are identified as physically equivalent — the more aggressively the unphysical degrees of freedom are removed from the theory. For a non-Abelian gauge theory such as $SU(3)_c$ QCD, the gauge orbit is so large that it removes not only the unphysical polarisations of the gluon (the longitudinal and temporal modes) but also generates a non-trivial interaction structure among the gluons themselves: the gluons carry colour charge and interact with each other, unlike the photon in the Abelian $U(1)$ case. This self-interaction of the gauge field is the mechanism behind *asymptotic freedom*: the effective coupling constant of QCD decreases at high energies, making the theory weakly coupled at short distances and strongly coupled at long distances. Asymptotic freedom — one of the most important properties of QCD, for which the 2004 Nobel Prize was awarded — is a direct consequence of the non-Abelian gauge structure, which is itself a consequence of the non-Abelian nature of the gauge orbit.

The force field is therefore not just a messenger that restores covariance: it is the agent of dimension reduction. The bigger the gauge group, the larger the orbit, the more dimensions are reduced, and the better the ultraviolet behaviour of the theory. The Standard Model’s gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$ was not chosen for aesthetic reasons: it is the gauge group that produces the observed forces, the observed particle content, and the observed ultraviolet behaviour of nature, all from the single principle that the physical degrees of freedom are gauge orbits, not fields.

The structure of this section. The remainder of this section makes these ideas precise. We derive the covariant derivative and the gauge field transformation law for a general gauge group, show how the field strength tensor and the gauge kinetic term are forced by the same covariance requirement, and write the complete locally gauge-invariant Lagrangian for the Standard Model. Throughout, the guiding principle is not symmetry but redundancy: every equation is an equation about gauge orbits, not about individual field configurations.

30.2 Local gauge invariance and its price

Now ask the question: what if we allow the phase to depend on spacetime position?

$$\phi(x) \longmapsto e^{i\alpha(x)} \phi(x), \quad \alpha(x) \text{ an arbitrary function of } x. \quad (30.1)$$

This is a *local* $U(1)$ transformation. The physical motivation is that the absolute phase of a quantum field at a single point is not observable—only phase *differences* between points are observable, and we should be free to redefine the phase independently at each point.

Check what happens to the kinetic term:

$$\begin{aligned} \partial_\mu \phi(x) &\longmapsto \partial_\mu (e^{i\alpha(x)} \phi(x)) \\ &= e^{i\alpha(x)} \partial_\mu \phi(x) + i(\partial_\mu \alpha(x)) e^{i\alpha(x)} \phi(x). \end{aligned} \quad (30.2)$$

The ordinary derivative ∂_μ is *not* covariant under local gauge transformations: it produces an extra term $i(\partial_\mu\alpha)e^{i\alpha}\phi$ that spoils the invariance of the Lagrangian. The kinetic term (29.4) is therefore *not* locally gauge invariant.

The extra term $\partial_\mu\alpha(x)$ is the problem. It arises because α varies from point to point, so comparing ϕ at two nearby points x and $x+dx$ requires knowing how the phase convention changes between them.

The fix: introduce a compensating vector field. We introduce a new field $A_\mu(x)$ —a vector field with one Lorentz index μ —and declare that under a local $U(1)$ transformation it transforms as:

$$A_\mu(x) \mapsto A_\mu(x) + \frac{1}{e} \partial_\mu\alpha(x). \quad (30.3)$$

Now define the *covariant derivative*:

$$D_\mu \equiv \partial_\mu - ieA_\mu(x). \quad (30.4)$$

Check how $D_\mu\phi$ transforms:

$$\begin{aligned} D_\mu\phi &\mapsto (\partial_\mu - ie(A_\mu + \tfrac{1}{e}\partial_\mu\alpha))e^{i\alpha}\phi \\ &= e^{i\alpha}\partial_\mu\phi + i(\partial_\mu\alpha)e^{i\alpha}\phi - ieA_\mu e^{i\alpha}\phi - i(\partial_\mu\alpha)e^{i\alpha}\phi \\ &= e^{i\alpha}(\partial_\mu - ieA_\mu)\phi \\ &= e^{i\alpha} D_\mu\phi. \end{aligned} \quad (30.5)$$

The two extra terms— $+i(\partial_\mu\alpha)e^{i\alpha}\phi$ from the derivative acting on $e^{i\alpha}$, and $-i(\partial_\mu\alpha)e^{i\alpha}\phi$ from the shift (30.3) of A_μ —cancel exactly. The covariant derivative $D_\mu\phi$ transforms in exactly the same way as ϕ itself: it picks up the factor $e^{i\alpha(x)}$ and nothing else.

This is the dance of the title: $\partial_\mu\alpha(x)$ is generated by the matter field's phase transformation, and A_μ shifts by exactly $\frac{1}{e}\partial_\mu\alpha(x)$ to cancel it. Matter generates a disturbance; the vector field steps in to compensate.

The locally gauge invariant Lagrangian. Replace every ordinary derivative ∂_μ with the covariant derivative D_μ , and add the kinetic term for A_μ :

$$\mathcal{L}_{\text{QED}} = (D^\mu\phi)^*(D_\mu\phi) - m^2\phi^*\phi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}, \quad (30.6)$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the field strength, which is itself gauge invariant:

$$F_{\mu\nu} \mapsto \partial_\mu(A_\nu + \tfrac{1}{e}\partial_\nu\alpha) - \partial_\nu(A_\mu + \tfrac{1}{e}\partial_\mu\alpha) = F_{\mu\nu} + \tfrac{1}{e}(\partial_\mu\partial_\nu - \partial_\nu\partial_\mu)\alpha = F_{\mu\nu}, \quad (30.7)$$

since partial derivatives commute. The Lagrangian (30.6) is invariant under the local $U(1)$ transformation (30.1) and (30.3).

Notice what we have *not* added: a mass term $\frac{1}{2}m_A^2 A^\mu A_\mu$ for the vector field. Such a term would transform as

$$A^\mu A_\mu \mapsto (A^\mu + \tfrac{1}{e}\partial^\mu\alpha)(A_\mu + \tfrac{1}{e}\partial_\mu\alpha) = A^\mu A_\mu + \frac{2}{e}A^\mu\partial_\mu\alpha + \dots, \quad (30.8)$$

which is not gauge invariant. Local gauge invariance therefore *forbids a mass term for the vector field*. The photon is massless not by accident but because it is the gauge boson of a local $U(1)$ symmetry.

The principle, stated cleanly.

Demanding invariance of the Lagrangian under local gauge transformations forces the existence of a vector field A_μ that couples to matter in a specific, completely determined way. The form of the coupling, the transformation law of A_μ , and the masslessness of A_μ are all consequences of the single requirement of local gauge invariance. We do not put the force in by hand; the force is the price of local symmetry.

This is why the Standard Model has the gauge bosons it does: each factor of the gauge group $U(1)_Y \times SU(2)_L \times SU(3)_c$ demands its own compensating vector field, and the properties of those vector fields are entirely determined by the symmetry.

30.3 Non-Abelian gauge invariance: $SU(N)$

The $U(1)$ story generalises to non-Abelian groups, but with one crucial new feature: the gauge bosons interact with *each other* as well as with matter.

Suppose the field $\Psi^i(x)$ carries an internal index $i \in \{1, \dots, N\}$ and transforms under a local $SU(N)$ transformation:

$$\Psi^i(x) \mapsto U(x)^i_{i'} \Psi^{i'}(x), \quad U(x) \in SU(N), \quad (30.9)$$

where $U(x)$ now depends on x . Writing $U(x) = e^{ig\theta^a(x)T^a}$ for generators T^a of $\mathfrak{su}(N)$ and real functions $\theta^a(x)$, the ordinary derivative again fails:

$$\partial_\mu \Psi^i \mapsto U^i_{i'} \partial_\mu \Psi^{i'} + (\partial_\mu U^i_{i'}) \Psi^{i'}. \quad (30.10)$$

The extra term $(\partial_\mu U)\Psi$ spoils invariance, just as $(\partial_\mu \alpha)e^{i\alpha}\phi$ did in the $U(1)$ case.

The non-Abelian covariant derivative. We introduce a vector field $A_\mu(x)$ that is now a *matrix* in the internal space:

$$A_\mu(x) = A_\mu^a(x) T^a \in \mathfrak{su}(N), \quad (A_\mu)^i_{i'} = A_\mu^a(x) (T^a)^i_{i'}, \quad (30.11)$$

where $a \in \{1, \dots, \dim G\}$ runs over the generators. The covariant derivative is:

$$(D_\mu \Psi)^i = \partial_\mu \Psi^i - ig (A_\mu)^i_j \Psi^j = \partial_\mu \Psi^i - ig A_\mu^a (T^a)^i_j \Psi^j. \quad (30.12)$$

The gauge field transforms to compensate:

$$A_\mu(x) \mapsto U(x) A_\mu(x) U(x)^{-1} + \frac{i}{g} U(x) \partial_\mu U(x)^{-1}. \quad (30.13)$$

The first term UAU^{-1} is a conjugation (the gauge field rotates in the adjoint representation); the second term $\frac{i}{g}U\partial_\mu U^{-1}$ is the compensating shift, generalising $\frac{1}{e}\partial_\mu \alpha$ from the $U(1)$ case. One can verify that $D_\mu \Psi \mapsto U(x) D_\mu \Psi$ under (30.9) and (30.13), so the covariant derivative transforms covariantly.

The non-Abelian field strength. The field strength for a non-Abelian gauge theory is:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu], \quad (30.14)$$

where $[A_\mu, A_\nu] = A_\mu A_\nu - A_\nu A_\mu$ is the matrix commutator. This extra commutator term has no $U(1)$ analogue (there $[A_\mu, A_\nu] = 0$ since numbers commute). It means $F_{\mu\nu}$ is *not* gauge invariant but transforms as $F_{\mu\nu} \mapsto U F_{\mu\nu} U^{-1}$; however, the kinetic term $\text{Tr}(F^{\mu\nu} F_{\mu\nu})$ is gauge invariant, since the trace is cyclic: $\text{Tr}(U F_{\mu\nu} U^{-1} U F^{\mu\nu} U^{-1}) = \text{Tr}(F_{\mu\nu} F^{\mu\nu})$.

Expanding $\text{Tr}(F^{\mu\nu} F_{\mu\nu})$ in components, the $[A, A]$ term in $F_{\mu\nu}$ generates cubic and quartic terms in A in the Lagrangian. These are *self-interaction vertices* of the gauge bosons: the gluons interact with each other, and the W bosons interact with each other. This has no analogue in QED (photons do not directly interact with photons at tree level), and it is the origin of quark confinement in QCD and of the rich phenomenology of the weak force.

30.4 The three gauge groups of the Standard Model

The Standard Model gauge group $G_{\text{SM}} = U(1)_Y \times SU(2)_L \times SU(3)_c$ contains three factors, each demanding its own compensating vector field. We now go through each in turn.

$U(1)_Y$: hypercharge and the origin of the photon.

The group is $U(1)_Y$, with gauge parameter $\alpha(x) \in \mathbb{R}$. Each field Ψ with hypercharge Y transforms as:

$$\Psi(x) \longmapsto e^{iY\alpha(x)} \Psi(x). \quad (30.15)$$

Demanding local invariance introduces one vector field $B_\mu(x)$, transforming as:

$$B_\mu(x) \longmapsto B_\mu(x) + \frac{1}{g'} \partial_\mu \alpha(x), \quad (30.16)$$

with covariant derivative:

$$D_\mu = \partial_\mu - ig' Y B_\mu. \quad (30.17)$$

The gauge boson B_μ is a single vector field with no internal index (the group is Abelian, dimension 1). After electroweak symmetry breaking, B_μ mixes with the neutral W_μ^3 boson of $SU(2)_L$ to produce the physical photon A_μ and the Z boson. The photon is the massless combination; the Z acquires mass from the Higgs mechanism.

$SU(2)_L$: weak isospin and the W bosons.

The group is $SU(2)_L$, acting only on left-chiral fields. A field Ψ_L^i with isospin index $i \in \{1, 2\}$ transforms as:

$$\Psi_L^i(x) \longmapsto V(x)^i_{i'} \Psi_L^{i'}(x), \quad V(x) = e^{ig\theta^a(x)\tau^a/2} \in SU(2), \quad (30.18)$$

where $\tau^a = \sigma^a$ are the Pauli matrices (generators of $\mathfrak{su}(2)$). Demanding local invariance introduces three vector fields $W_\mu^a(x)$, $a \in \{1, 2, 3\}$ (one per generator of $SU(2)$), packaged as a matrix:

$$W_\mu(x) = W_\mu^a(x) \frac{\tau^a}{2} = \frac{1}{2} \begin{pmatrix} W_\mu^3 & W_\mu^1 - iW_\mu^2 \\ W_\mu^1 + iW_\mu^2 & -W_\mu^3 \end{pmatrix}. \quad (30.19)$$

The covariant derivative acting on a left-chiral isospin doublet is:

$$(D_\mu \Psi_L)^i = \partial_\mu \Psi_L^i - ig(W_\mu)^i_{i'} \Psi_L^{i'} = \partial_\mu \Psi_L^i - ig W_\mu^a \left(\frac{\tau^a}{2}\right)^i_{i'} \Psi_L^{i'}. \quad (30.20)$$

The physical $W^\pm$ bosons are the combinations $W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2)$, and W_μ^3 mixes with B_μ to give the Z and the photon.

Because $SU(2)_L$ acts *only on left-chiral fields*, the covariant derivative D_μ acts differently on Ψ_L and Ψ_R :

$$D_\mu \Psi_L^i = \partial_\mu \Psi_L^i - ig W_\mu^a (\tau^a/2)^i_{i'} \Psi_L^{i'} - ig' Y_L B_\mu \Psi_L^i, \quad (30.21)$$

$$D_\mu \Psi_R = \partial_\mu \Psi_R - ig' Y_R B_\mu \Psi_R. \quad (30.22)$$

The W boson term is absent from the right-chiral derivative. The force field knows about chirality because the symmetry group $SU(2)_L$ was defined to act only on left-chiral fields. This is the precise sense in which the W bosons “see” only left-handed particles.

$SU(3)_c$: colour and the gluons.

The group is $SU(3)_c$, acting on the colour index $\alpha \in \{r, g, b\}$ of quarks. A quark field $q^\alpha(x)$ transforms as:

$$q^\alpha(x) \longmapsto U(x)^\alpha_{\alpha'} q^{\alpha'}(x), \quad U(x) = e^{ig_s \theta^a(x) T^a} \in SU(3), \quad (30.23)$$

where $T^a = \lambda^a/2$ ($a = 1, \dots, 8$) are the Gell-Mann matrices, generators of $\mathfrak{su}(3)$. Demanding local invariance introduces eight gluon fields $G_\mu^a(x)$, $a \in \{1, \dots, 8\}$:

$$G_\mu(x) = G_\mu^a(x) T^a \in \mathfrak{su}(3), \quad (G_\mu)^\alpha_{\alpha'} = G_\mu^a (T^a)^\alpha_{\alpha'}. \quad (30.24)$$

The covariant derivative acting on a quark is:

$$(D_\mu q)^\alpha = \partial_\mu q^\alpha - ig_s (G_\mu)^\alpha_{\alpha'} q^{\alpha'} = \partial_\mu q^\alpha - ig_s G_\mu^a (T^a)^\alpha_{\alpha'} q^{\alpha'}. \quad (30.25)$$

The non-Abelian field strength is:

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c, \quad (30.26)$$

where f^{abc} are the $SU(3)$ structure constants defined by $[T^a, T^b] = if^{abc}T^c$. The $f^{abc}G^bG^c$ term gives rise to three-gluon and four-gluon vertices in the Lagrangian. Gluons carry colour charge and therefore couple to each other—unlike photons, which are electrically neutral and do not couple directly to each other. This gluon self-coupling is responsible for asymptotic freedom (the force weakens at short distances) and confinement (quarks cannot be isolated).

The complete Standard Model Lagrangian structure. The full Lagrangian is:

$$\mathcal{L}_{\text{SM}} = \underbrace{-\frac{1}{4}B^{\mu\nu}B_{\mu\nu} - \frac{1}{4}W^{a,\mu\nu}W_{\mu\nu}^a - \frac{1}{4}G^{a,\mu\nu}G_{\mu\nu}^a}_{\text{gauge kinetic terms}} + \underbrace{\bar{\Psi} i\gamma^\mu D_\mu \Psi}_{\text{matter kinetic + coupling}} + \underbrace{|D_\mu \phi|^2 - V(\phi)}_{\text{Higgs}} + \underbrace{y \bar{\Psi}_L \phi \Psi_R + \text{h.c.}}_{\text{Yukawa (masses)}}. \quad (30.27)$$

Every term is a Lorentz scalar (all μ, ν and spinor indices contracted), every term is invariant under $U(1)_Y \times SU(2)_L \times SU(3)_c$ (all internal indices contracted or paired under gauge transformations), and every term has mass dimension ≤ 4 . The covariant derivative D_μ in the matter term packages all three gauge couplings simultaneously:

$$D_\mu = \partial_\mu - ig'YB_\mu - igW_\mu^a T_{SU(2)}^a - ig_s G_\mu^a T_{SU(3)}^a. \quad (30.28)$$

Each term acts on the appropriate internal index: B_μ multiplies by a phase (acting on the $U(1)$ charge), $W_\mu^a T^a$ acts on the isospin index i , and $G_\mu^a T^a$ acts on the colour index α . The three forces are three independent matrix actions on three independent index slots of the same field tensor—exactly the structure we built up through this chapter.

Remark. [Force from symmetry: the summary] The logical chain is:

$$\underbrace{\text{internal index on matter field}}_{\text{e.g., colour index } \alpha} \xrightarrow{\text{demand local invariance}} \underbrace{\text{compensating vector field}}_{\text{e.g., gluon } G_\mu^a} \xrightarrow{\text{add kinetic term}} \underbrace{\text{force with determined coupling}}_{\text{e.g., QCD}}.$$

The matter field carries an index. Local gauge invariance demands a vector field to compensate phase variations between points. The kinetic term for the vector field gives it dynamics. The result is a force whose strength, coupling structure, and gauge boson properties are all fixed by the symmetry group alone. We do not choose the form of the force; we choose the symmetry, and the force follows.

30.5 The adjoint representation: gauge bosons as Lie algebra elements

We have established that demanding local gauge invariance under a group G forces the existence of a compensating vector field. We now ask a sharper question: what representation of G does this vector field carry? The answer is always the same, regardless of the gauge group: *gauge bosons transform in the adjoint representation*. This single fact explains why there is one photon, three W bosons, and eight gluons.

How many gauge bosons does a group require?

The gauge field $A_\mu(x) = A_\mu^a(x)T^a$ is a *Lie-algebra-valued 1-form*: at each spacetime point x and for each direction μ , it takes a value in the Lie algebra $\mathfrak{g}$ of the gauge group G . The Lie algebra has one basis element T^a per generator of G , so the number of independent components of A_μ per spacetime point equals $\dim G$, the dimension of G as a Lie group.

Under a local gauge transformation $U(x) = e^{i\alpha^a(x)T^a}$, the gauge field must absorb the $\dim G$ independent functions $\alpha^a(x)$: one gauge field component per generator. This fixes the count completely:

Group	$\dim G$	Gauge bosons	Physical fields
$U(1)_Y$	1	B_μ	1
$SU(2)_L$	3	$W_\mu^a, a \in \{1, 2, 3\}$	3
$SU(3)_c$	8	$G_\mu^a, a \in \{1, \dots, 8\}$	8

The dimension of each gauge group follows from a simple parameter count. An $N \times N$ complex matrix has $2N^2$ real parameters; unitarity $U^\dagger U = I$ imposes N^2 real constraints; $\det U = 1$ imposes one more. So $\dim SU(N) = N^2 - 1$. For $N = 2$: $\dim SU(2) = 3$. For $N = 3$: $\dim SU(3) = 8$. For $U(1)$: $\dim U(1) = 1$ (one phase parameter).

The adjoint representation

The *adjoint representation* is the action of G on its own Lie algebra $\mathfrak{g}$ by conjugation. For a group element $U \in G$ and a Lie algebra element $X = X^a T^a \in \mathfrak{g}$:

$$X \mapsto U X U^\dagger = (Ad_U)^a_b X^b T^a, \quad (30.29)$$

where $(Ad_U)^a_b$ is the $(\dim G) \times (\dim G)$ matrix of the adjoint representation. For infinitesimal $U = I + i\alpha^c T^c$, using the Lie algebra commutation relations $[T^a, T^b] = if^{abc}T^c$:

$$(Ad_U)^a_b = \delta_b^a - \alpha^c f^{cab} + O(\alpha^2), \quad (30.30)$$

so the adjoint representation matrices are built directly from the structure constants f^{abc} .

The gauge field $A_\mu = A_\mu^a T^a$ transforms under a gauge transformation as:

$$A_\mu \mapsto U A_\mu U^\dagger + \frac{i}{g}(\partial_\mu U)U^\dagger. \quad (30.31)$$

The first term $U A_\mu U^\dagger$ is precisely the adjoint action (30.29) on the Lie-algebra component of A_μ . The second term $\frac{i}{g}(\partial_\mu U)U^\dagger$ is the inhomogeneous shift that compensates the matter field's phase variation — the same shift we saw in the covariant derivative construction.

This is why gauge bosons are always in the adjoint representation: the gauge field *is* a Lie algebra element, and the Lie algebra transforms under G by conjugation, which is the adjoint action by definition. Quarks are acted upon by the gauge group; gluons *are* the gauge transformation, and the group acts on itself by conjugation.

The three gauge groups: adjoint representations explicitly

$U(1)_Y$: one B boson, trivial adjoint.

$U(1)$ is Abelian: all elements commute, so $UXU^\dagger = X$ for all X . The structure constants are all zero: $f^{abc} = 0$. The adjoint representation is trivial (the identity map): the B_μ field does not transform under the adjoint action. The inhomogeneous shift $\frac{1}{g}\partial_\mu \alpha$ in (30.31) is the entire transformation.

Consequence: B_μ carries no hypercharge. It does not interact with other B_μ bosons at tree level. There are no $B_\mu B_\mu B_\mu$ vertices in the Lagrangian.

$SU(2)_L$: three W bosons, adjoint of $SU(2)$.

$SU(2)$ has generators $T^a = \sigma^a/2$ ($a \in \{1, 2, 3\}$) with structure constants $f^{abc} = \varepsilon^{abc}$ (the totally antisymmetric Levi-Civita symbol). The adjoint representation is 3×3 (dimension $\dim SU(2) = 3$), acting on the three gauge fields W_μ^a :

$$W_\mu^a \mapsto (Ad_V)^a_b W_\mu^b \quad (\text{homogeneous part}), \quad (Ad_V)^a_b = \delta_b^a - \alpha^c \varepsilon^{cab} + \dots \quad (30.32)$$

The W bosons carry weak isospin. Expanding the kinetic term $-\frac{1}{4}W^{a,\mu\nu}W_{\mu\nu}^a$ with

$$W_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g \varepsilon^{abc} W_\mu^b W_\nu^c, \quad (30.33)$$

the $\varepsilon^{abc}W^b W^c$ term generates cubic and quartic W self-interaction vertices. The physical $W^\pm$ bosons and the neutral W^3 interact with each other, as observed in W^+W^- pair production at the LHC.

After electroweak symmetry breaking, W_μ^3 mixes with B_μ to produce the physical Z boson and the photon:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos \theta_W & \sin \theta_W \\ -\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix}, \quad (30.34)$$

where θ_W is the Weinberg angle. The photon A_μ is the massless combination; the Z acquires a mass $m_Z = m_W / \cos \theta_W$ from the Higgs mechanism.

$SU(3)_c$: eight gluons, adjoint of $SU(3)$.

$SU(3)$ has 8 generators $T^a = \lambda^a/2$ (the Gell-Mann matrices) with structure constants f^{abc} defined by $[T^a, T^b] = i f^{abc} T^c$. The adjoint representation is 8×8 , acting on the eight gluon fields G_μ^a :

$$G_\mu^a \mapsto (Ad_U)^a_b G_\mu^b \quad (\text{homogeneous part}). \quad (30.35)$$

The field strength:

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c \quad (30.36)$$

generates three-gluon and four-gluon self-interaction vertices. Because the f^{abc} are non-zero for $SU(3)$ (the group is non-Abelian), gluons carry colour charge and interact with each other. This is the mechanism behind asymptotic freedom and confinement.

Fundamental vs. adjoint: the complete index picture

We now have two kinds of colour-like indices: the *fundamental* index (carried by matter fields, acted on by U^a_b) and the *adjoint* index (carried by gauge bosons, acted on by $(Ad_U)^a_b$).

Field	Group	Rep	Index range	Transform
B_μ	$U(1)$	trivial adjoint	none	phase shift only
W_μ^a	$SU(2)_L$	adjoint 3	$a \in \{1, 2, 3\}$	$(Ad_V)^a_b W_\mu^b$
G_μ^a	$SU(3)_c$	adjoint 8	$a \in \{1, \dots, 8\}$	$(Ad_U)^a_b G_\mu^b$
lepton L_L^i	$SU(2)_L$	fundamental 2	$i \in \{1, 2\}$	$V^i_{i'} L_L^{i'}$
quark q^α	$SU(3)_c$	fundamental 3	$\alpha \in \{r, g, b\}$	$U^\alpha_{\alpha'} q^{\alpha'}$
antiquark $\bar{q}_\alpha$	$SU(3)_c$	anti-fundamental $\bar{3}$	$\alpha \in \{r, g, b\}$	$(U^\dagger)_\alpha^{\alpha'} \bar{q}_{\alpha'}$
Higgs ϕ^i	$SU(2)_L$	fundamental 2	$i \in \{1, 2\}$	$V^i_{i'} \phi^{i'}$

The key distinction: matter fields carry *fundamental* indices because they are *acted on* by the gauge group. Gauge bosons carry *adjoint* indices because they *are* the gauge transformation, and the group acts on itself by conjugation. This is not a choice—it follows necessarily from the structure of the covariant derivative and the transformation law (30.31).

Self-interaction and the adjoint. A gauge boson carries a non-trivial adjoint index if and only if the gauge group is non-Abelian. For $U(1)$ the adjoint is trivial: the photon B_μ carries no charge and has no self-interactions. For $SU(2)_L$ and $SU(3)_c$ the adjoint is non-trivial: the W bosons and gluons carry charge under their own gauge group and interact with each other. The self-interaction coupling is always proportional to the structure constants f^{abc} , which measure the failure of the generators to commute. Commutativity ($f^{abc} = 0$, Abelian) means no self-interaction; non-commutativity means the messengers message each other.

Remark. [ML analogy: weight matrices vs. activations] In a neural network, activations (matter fields) are acted on by weight matrices (gauge group elements). The weight matrices themselves are not activations—they live in a different space (the space of linear maps, analogous to the Lie algebra) and transform differently when the network architecture is changed. The adjoint representation is the transformation law of the “weight matrices” (gauge bosons) under the symmetry group, as opposed to the transformation law of the “activations” (matter fields). The distinction between fundamental and adjoint is the distinction between data and the operators that act on data.

30.6 Gauge orbits: physical states as equivalence classes

The gauge transformation $A_\mu \rightarrow A_\mu + \frac{1}{e}\partial_\mu\alpha$ reveals something important: the gauge field $A_\mu(x)$ is not itself a physical observable. Two field configurations related by such a transformation give identical electric and magnetic fields, identical particle trajectories, and identical scattering amplitudes. They are two descriptions of the same physical state.

The *gauge orbit* of a configuration A_μ is the set of all configurations physically identical to it:

$$[A_\mu] = \left\{ A_\mu + \frac{1}{e}\partial_\mu\alpha : \alpha(x) \in C^\infty \right\}, \quad (30.37)$$

and for $SU(N)$:

$$[A_\mu] = \left\{ U A_\mu U^\dagger + \frac{i}{g}(\partial_\mu U)U^\dagger : U(x) \in SU(N) \right\}. \quad (30.38)$$

The physical configuration space is not the space of all A_μ , but the *quotient*:

$$\mathcal{A}_{\text{phys}} = \mathcal{A}/\mathcal{G}, \quad (30.39)$$

where $\mathcal{G}$ is the group of all gauge transformations. Each point in $\mathcal{A}_{\text{phys}}$ is an entire orbit—a collection of field configurations that are all physically the same.

The true gauge-invariant content of A_μ is captured by the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, which encodes the electric field $E_i = F_{0i}$ and magnetic field $B_i = \epsilon_{ijk}F_{jk}$. These are the actual observables. The potential A_μ is a convenient but redundant intermediary.

Gauge fixing. To do calculations, one picks one representative configuration from each orbit—a *gauge choice*. Common choices include:

- *Lorenz gauge:* $\partial^\mu A_\mu = 0$ (manifestly Lorentz covariant, useful for covariant Feynman rules).
- *Coulomb gauge:* $\nabla \cdot \mathbf{A} = 0$ (makes the two physical photon polarisations explicit).
- *Axial gauge:* $A_3 = 0$ (removes ghost complications in some calculations).

No physical prediction depends on the gauge choice; gauge fixing is purely a computational convenience. The Faddeev–Popov procedure implements this quotient systematically in the path integral.

Gauge theory	ML / mathematics
Gauge field $A_\mu(x)$	Redundant parametrisation
Gauge transformation	Group action on parameter space
Gauge orbit $[A_\mu]$	Equivalence class under group action
Physical space $\mathcal{A}/\mathcal{G}$	Quotient space
Gauge fixing $\partial^\mu A_\mu = 0$	Choosing a canonical representative
Field strength $F_{\mu\nu}$	Gauge-invariant observable

30.7 Gauge invariance as dimension reduction

We asked earlier *why* the Lagrangian must be locally gauge invariant, and answered: because the absolute phase of a quantum field at a single point is unobservable. Here we give a complementary answer from the perspective of degrees of freedom: *gauge invariance is a dimension reduction that removes unphysical modes and keeps the dynamics stable*.

Counting degrees of freedom. A photon field A_μ has four components per spacetime point. Naively this suggests four propagating degrees of freedom. But:

- The $U(1)$ gauge orbit removes one degree of freedom (one function $\alpha(x)$).
- The transversality constraint $k^\mu \epsilon_\mu = 0$ (which follows from the equations of motion) removes one more.

The result is $4 - 2 = 2$ physical degrees of freedom: the two transverse helicities of the photon. The four-component field A_μ is a redundant description of a two-component physical object.

For the gluon field G_μ^a of $SU(3)_c$: there are $4 \times 8 = 32$ components per spacetime point. The gauge orbit removes 8 dimensions (one per generator) and transversality removes 8 more, leaving $32 - 16 = 16 = 8 \times 2$ physical degrees of freedom: two helicities for each of the eight gluons. The general pattern is: a gauge group of dimension $\dim G$ removes $2 \dim G$ degrees of freedom per momentum mode.

Why this matters: ghost modes. Without gauge invariance, all components of A_μ would propagate as independent degrees of freedom, including the timelike component A_0 . In the Lagrangian, this component has a kinetic term with the *wrong sign* (a ghost mode), giving a Hamiltonian unbounded below. The Schrödinger evolution $\dot{h}_t = -iHh_t$ would be unstable: the Fock hidden state would acquire unbounded amplitudes in the ghost directions.

Gauge invariance removes the ghost from the physical Hilbert space by declaring it unphysical. The physical Hamiltonian, restricted to gauge-invariant states, is positive-definite and the evolution is stable. Gauge invariance is the mechanism that prevents the dynamics from crashing.

Non-Abelian gauge groups and asymptotic freedom. The dimension reduction is more aggressive for larger gauge groups. This has a striking consequence: in QCD with gauge group $SU(3)_c$, the gluons carry colour charge and interact with each other via the f^{abc} structure constants. These self-interactions contribute to the running of the coupling in the *opposite* direction from the matter fields, giving *asymptotic freedom*: the coupling $\alpha_s(\mu)$ decreases at high energies and grows at low energies. This is why quarks behave as nearly free particles at short

distances but are permanently confined at long distances. The non-Abelian orbit structure is directly responsible for confinement.

In short: matter fields are the data; force fields are the regularizers; gauge invariance is the constraint that keeps the computation from diverging.

30.8 The messenger picture: what gauge bosons actually carry

Popular accounts describe forces as exchange of messenger particles: electrons repel by exchanging photons, quarks attract by exchanging gluons. This picture is correct as far as it goes, but it misses something. The gauge boson is not just carrying momentum—it is carrying a *reference frame*.

The comparison problem. The matter field $\psi(x)$ at spacetime point x carries an internal phase (a colour vector, an isospin component) defined at x . The field $\psi(y)$ at a different point y carries its own internal phase, defined independently. Because gauge transformations can act differently at x and y , there is no canonical way to compare the two phases directly: the comparison depends on the path taken from x to y .

This is the address-book problem: two people at different locations want to compare notes, but each has their own local coordinate system. A messenger must physically carry the translation between coordinate systems from one location to the other.

The covariant derivative as parallel transport. The ordinary derivative $\partial_\mu\psi$ compares $\psi(x+dx)$ with $\psi(x)$ directly. But these fields live at different points where different gauge transformations can act—the comparison is frame-dependent and therefore unphysical. The covariant derivative corrects for this:

$$D_\mu\psi(x) = \lim_{dx \rightarrow 0} \frac{U(x+dx, x) \psi(x+dx) - \psi(x)}{dx}, \quad (30.40)$$

where $U(x+dx, x) = I - igA_\mu(x)dx$ is the *parallel transporter*: the infinitesimal gauge transformation that carries the internal reference frame from $x+dx$ back to x so that the two field values can be compared in the same frame. Expanding:

$$D_\mu\psi = (\partial_\mu - igA_\mu)\psi. \quad (30.41)$$

The term $-igA_\mu\psi$ is the messenger’s correction: it adjusts the raw derivative by the amount the internal reference frame has rotated between x and $x+dx$. The gauge field A_μ is a *connection*—a mathematical object specifying how to consistently compare field values at neighbouring points.

What the photon propagator carries. At lowest order in QED, two electrons interact by exchanging a photon. The amplitude is:

$$i\mathcal{M} = (-ie)^2 [\bar{u}(p')\gamma^\mu u(p)] \frac{-i\eta_{\mu\nu}}{k^2} [\bar{u}(q')\gamma^\nu u(q)], \quad (30.42)$$

where $k^\mu = p - p'$ is the transferred 4-momentum. The photon propagator $-i\eta_{\mu\nu}/k^2$ carries two things simultaneously:

- **4-momentum** k^μ : the photon transfers energy and momentum from one electron to the other. This is the “ball” in the exchange-particle metaphor.
- **A Lorentz index** μ : the photon carries a polarisation index that contracts with the electron’s current $\bar{u}\gamma^\mu u$ at each vertex. This index encodes the orientation of the interaction—which polarisation component of the field is coupling to which component of the current. This is the reference frame being transported.

In the non-relativistic limit, (30.42) reproduces the Coulomb potential $V(r) = e^2/4\pi r$.

Non-Abelian messengers talk to each other. In QED the photon is electrically neutral: it is a messenger that does not message itself. In QCD the gluons carry colour charge. The field strength:

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c \quad (30.43)$$

contains the term $g_s f^{abc} G_\mu^b G_\nu^c$, generating three-gluon and four-gluon self-interaction vertices in the Lagrangian. The structure constant f^{abc} (defined by $[T^a, T^b] = i f^{abc} T^c$) measures how much the generators fail to commute: the messengers interact with each other in direct proportion to the non-commutativity of the gauge group.

This self-coupling is the reason QCD confines quarks. In QED, photon messengers fly between charges independently. In QCD, gluon messengers interact with each other, concentrating the colour flux into a narrow tube rather than spreading it as a Coulomb field. The force between quarks therefore grows with distance ($V(r) \sim \sigma r$ at large r), making quark isolation impossible.

Metaphor	Precise content
Houses at different addresses	Matter fields $\psi(x)$, $\psi(y)$ at different spacetime points
Each house has its own coordinates	Local gauge symmetry: independent phase at each x
Messenger carries coordinate translation	Gauge field A_μ = connection (parallel transporter)
Messenger needed to compare notes	Ordinary ∂_μ fails; covariant D_μ required
Exchanging a ball	Photon propagator $-i\eta_{\mu\nu}/k^2$ transfers 4-momentum
Coulomb repulsion	Non-relativistic limit of (30.42): $V(r) = e^2/4\pi r$
Messenger talks to itself	Gluon self-coupling $g_s f^{abc} G_\mu^b G_\nu^c$ in (30.43)
Messenger gets tangled	Colour flux tube, linear potential, confinement

The exchange-particle metaphor captures momentum transfer and Coulomb repulsion correctly. What it misses is that the messenger is also carrying a reference frame—the information needed to consistently compare quantum phases at different spacetime points. The force is not primitive; it is the consequence of maintaining consistent comparison across a space where the phase convention can vary freely from point to point.

From field to Fock vector

The indices on the classical field $\Psi^{m\alpha i}(x)$ become the labels on the Fock basis vectors after quantization. The mode expansion replaces the spacetime coordinate x with the momentum $\mathbf{k}$ via the Fourier transform for each time t :

$$x \longleftrightarrow \mathbf{k} \quad (\text{Fourier duality: position basis and momentum basis}), \quad (30.44)$$

where $\mathbf{k}$ is spatial momentum. The creation operator $a^{\dagger m\alpha i}(\mathbf{k})$ adds to the Fock hidden state one particle with Lorentz index m , colour α , isospin i , and momentum $\mathbf{k}$. The Fock basis vector:

$$|n; (\mathbf{k}_1, m_1, \alpha_1, i_1, f_1, \pm_1), \dots, (\mathbf{k}_n, m_n, \alpha_n, i_n, f_n, \pm_n)\rangle \quad (30.45)$$

is labelled by the index tuple for each of the n particles. The labels inside the ket come directly from the index slots of the classical field tensor, augmented by the particle/antiparticle label $\pm$ and the flavour label f .

The Hamiltonian couples the index slots. After quantization, the Hamiltonian H_{SM} is a polynomial in the creation and annihilation operators $a^{\dagger mai}(\mathbf{k})$ and $a^{mai}(\mathbf{k})$. Each interaction vertex (QED, weak, QCD, Yukawa) contracts specific index slots between different fields, in a way determined by the gauge invariance requirement (all indices must form singlets). The Schrödinger equation $\dot{h}_t = -iH_{\text{SM}} h_t$ rotates the Fock hidden state (30.45) among all basis vectors, with the interaction vertices driving amplitude flow between different particle configurations.

The story told by this rotation is the Standard Model: quarks exchange gluons (colour index contractions), electrons emit photons (charge contractions), left-handed fermion doublets interact via W bosons (isospin index contractions), and the Higgs couples left and right chiralities to generate masses (contracting the spinor index m between left and right Weyl components). The actors—each carrying their full costume of indices—interact through vertices that respect all the index structure simultaneously.

Throughout, we keep the Fock hidden state in view. The basis vectors of $\mathcal{F}_{\text{SM}}$ are labelled by all quantum numbers simultaneously—Lorentz spin, colour, isospin, flavour, momentum. Adding a new index to a field adds a new label to the Fock basis vectors. The Schrödinger RNN $\dot{h}_t = -iH_{\text{SM}} h_t$ runs on this larger labelled space. The index structure is the structure of the labels; the RNN is unchanged.

30.9 Yang and Mills (1954): from $U(1)$ to $SU(2)$

Maxwell’s electromagnetism is a gauge theory with gauge group $U(1)$: the simplest possible. The gauge field A_μ is a single real 4-vector, the gauge transformation is $A_\mu \rightarrow A_\mu + \partial_\mu \chi$ (an additive shift by a gradient), and the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ involves only derivatives of A_μ . Everything is linear: the field strength is linear in A_μ , the gauge transformation is linear, and the gauge field does not interact with itself.

The reason is that $U(1)$ is *Abelian*: its elements commute. The phase rotations $e^{i\alpha}$ and $e^{i\beta}$ satisfy $e^{i\alpha}e^{i\beta} = e^{i\beta}e^{i\alpha}$: the order does not matter. The commutator $[e^{i\alpha}, e^{i\beta}] = 0$ vanishes identically. This commutativity is why the $U(1)$ gauge field has no self-interaction: there is no commutator term to generate one.

In 1954, Chen-Ning Yang and Robert Mills asked: what happens when the gauge group is non-Abelian?

The problem: local $SU(2)$ invariance

Yang and Mills considered a matter field ψ transforming as a doublet under an internal $SU(2)$ symmetry (which they identified with isospin: the symmetry between proton and neutron). The global symmetry transformation is:

$$\psi(x) \rightarrow U \psi(x), \quad U \in SU(2), \quad U^\dagger U = \mathbf{1}, \quad \det U = 1. \quad (30.46)$$

The Lagrangian $\mathcal{L} = \bar{\psi} i \gamma^\mu \partial_\mu \psi$ is invariant under this global transformation because U is constant—it does not depend on x .

Now promote the symmetry to a *local* one: allow U to depend on the spacetime point:

$$\psi(x) \rightarrow U(x) \psi(x), \quad U(x) \in SU(2) \text{ for each } x. \quad (30.47)$$

The derivative $\partial_\mu \psi$ is no longer covariant: it picks up an extra term $(\partial_\mu U) \psi$ from the x -dependence of U . To restore invariance, a gauge field must be introduced—the Yang–Mills field B_μ .

The $U(1)$ case: why the commutator disappears

For $U(1)$, the gauge transformation of the connection is:

$$A_\mu \rightarrow A_\mu + \frac{1}{e} \partial_\mu \alpha, \quad (30.48)$$

where $U(x) = e^{i\alpha(x)}$ and $\alpha(x)$ is a real function. The field strength:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (30.49)$$

is gauge-invariant because $\partial_\mu \partial_\nu \alpha = \partial_\nu \partial_\mu \alpha$: the mixed partial derivatives commute. There is no commutator term in $F_{\mu\nu}$ because $U(1)$ is Abelian—the gauge parameters at different points commute, and the would-be commutator $[A_\mu, A_\nu]$ vanishes because A_μ is a single real number at each point, and real numbers commute.

The $SU(2)$ case: the commutator appears

For $SU(2)$, the gauge field B_μ is a 2×2 traceless Hermitian matrix at each spacetime point—equivalently, a triplet of real fields B_μ^a ($a = 1, 2, 3$) multiplied by the $SU(2)$ generators $\tau^a = \sigma^a/2$:

$$B_\mu(x) = B_\mu^a(x) \tau^a. \quad (30.50)$$

The covariant derivative is:

$$D_\mu \psi = \partial_\mu \psi - ig B_\mu \psi, \quad (30.51)$$

where g is the gauge coupling. Under the local transformation (30.47), the gauge field must transform as:

$$B_\mu \rightarrow U B_\mu U^\dagger + \frac{i}{g} (\partial_\mu U) U^\dagger \quad (30.52)$$

to make $D_\mu \psi \rightarrow U D_\mu \psi$ (covariant).

Now compute the field strength. In the $U(1)$ case, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is gauge-invariant. Try the same thing for $SU(2)$:

$$\partial_\mu B_\nu - \partial_\nu B_\mu \xrightarrow{\text{gauge}} U(\partial_\mu B_\nu - \partial_\nu B_\mu)U^\dagger + \text{extra terms}. \quad (30.53)$$

The extra terms do not cancel: $\partial_\mu B_\nu - \partial_\nu B_\mu$ is *not* gauge-covariant for $SU(2)$. The reason is that B_μ is a matrix, and the gauge transformation (30.52) involves $U B_\mu U^\dagger$ (conjugation), not $B_\mu + \text{gradient}$ (addition). Conjugation generates cross terms that the $U(1)$ additive transformation does not.

Yang and Mills' key insight: the extra terms can be cancelled by adding the *commutator* $[B_\mu, B_\nu]$ to the field strength. The gauge-covariant field strength is:

$$\boxed{G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu - ig[B_\mu, B_\nu]}. \quad (30.54)$$

Under the gauge transformation:

$$G_{\mu\nu} \rightarrow U G_{\mu\nu} U^\dagger. \quad (30.55)$$

The field strength transforms *covariantly* (by conjugation), not invariantly. The Lagrangian $\text{Tr}[G_{\mu\nu} G^{\mu\nu}]$ is gauge-invariant because the trace is cyclic: $\text{Tr}[UGU^\dagger UGU^\dagger] = \text{Tr}[G^2]$.

Why the commutator term is necessary. The commutator $[B_\mu, B_\nu] = B_\mu B_\nu - B_\nu B_\mu$ is nonzero because B_μ and B_ν are matrices that do not commute. In components:

$$[B_\mu, B_\nu] = B_\mu^a B_\nu^b [\tau^a, \tau^b] = i B_\mu^a B_\nu^b \epsilon^{abc} \tau^c. \quad (30.56)$$

The structure constants ϵ^{abc} of $SU(2)$ appear because the generators τ^a do not commute: $[\tau^a, \tau^b] = i\epsilon^{abc}\tau^c$. For $U(1)$, there is one generator ($\tau = 1$) and $[1, 1] = 0$: the commutator vanishes and (30.54) reduces to $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$.

The commutator term $-ig[B_\mu, B_\nu]$ in the field strength is what makes non-Abelian gauge theories qualitatively different from electromagnetism. It means:

- The field strength $G_{\mu\nu}$ is *nonlinear* in the gauge field B_μ : it contains a term quadratic in B .
- The kinetic Lagrangian $\text{Tr}[G_{\mu\nu}G^{\mu\nu}]$, when expanded, contains three- B and four- B vertices: the gauge field interacts with itself.
- The photon does not self-interact; the gluon does. This self-interaction is the origin of asymptotic freedom, confinement, and the qualitative differences between QED and QCD.

The Yang–Mills Lagrangian. The full Yang–Mills Lagrangian for the gauge field coupled to a fermion doublet is:

$$\mathcal{L}_{\text{YM}} = -\frac{1}{2}\text{Tr}[G_{\mu\nu}G^{\mu\nu}] + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi, \quad (30.57)$$

where $D_\mu = \partial_\mu - igB_\mu$ is the covariant derivative. The first term is the gauge field kinetic term (including the self-interaction from the commutator); the second is the matter kinetic term with the gauge coupling.

Pauli’s objection: where is the mass?

Yang presented this theory at a seminar at the Institute for Advanced Study in Princeton in February 1954. In the audience was Wolfgang Pauli, who had independently considered similar ideas but had not published them.

Pauli immediately identified the critical problem. He interrupted Yang’s talk repeatedly with the same question: “What is the mass of this field B_μ ?”

Yang’s theory required the gauge field B_μ to be massless, for the same reason the photon is massless: a mass term $m_B^2 \text{Tr}[B_\mu B^\mu]$ is not gauge-invariant under (30.52). But the force carriers of the nuclear force (the application Yang had in mind) must be massive: the nuclear force has a finite range ~ 1 fm, which requires massive mediators. A massless $SU(2)$ gauge boson would produce a long-range force like electromagnetism—which is not observed.

Pauli’s question was precise and devastating: the theory is mathematically beautiful, but physically it predicts massless gauge bosons that do not exist. Yang had no answer. According to accounts of the talk, Yang became so flustered by Pauli’s persistent questioning that he sat down before finishing. J. Robert Oppenheimer, who was chairing the session, intervened to let Yang continue.

Why Pauli did not publish. Pauli had arrived at the same non-Abelian gauge structure independently in private notes, but he did not publish precisely because of the mass problem. He considered the theory incomplete without a mechanism for giving mass to the gauge bosons. Yang and Mills published despite the problem, presenting the mathematical structure and leaving the mass question open.

The resolution: two decades away

Pauli’s question would not be answered for nearly two decades, and the answer required two separate breakthroughs:

The Higgs mechanism (1964). Englert, Brout, Higgs, Guralnik, Hagen, and Kibble showed that a scalar field with a Mexican-hat potential, coupled to the gauge field, can give mass to the gauge bosons while preserving gauge invariance of the Lagrangian. The gauge bosons acquire mass not from an explicit mass term $m^2 B_\mu B^\mu$ (which is forbidden) but from the vacuum expectation value of the scalar field: $m_W = gv/2$. The mass is a consequence of spontaneous symmetry breaking, not of an explicit breaking of gauge invariance.

Renormalizability ('t Hooft, 1971). Even with the Higgs mechanism, it was not obvious that the resulting theory—a spontaneously broken non-Abelian gauge theory—was renormalizable. Gerard 't Hooft, a graduate student of Martinus Veltman, proved in 1971 that it is: the divergent part of $\log Z$ can be absorbed into finitely many counterterms, preserving the BRST structure at every loop order. This was the result that made the electroweak theory ($SU(2)_L \times U(1)_Y$, spontaneously broken to $U(1)_{\text{em}}$) a calculable, predictive quantum field theory. 't Hooft and Veltman received the Nobel Prize in 1999.

The Standard Model (Glashow, Weinberg, Salam, 1961–68). With the Yang–Mills structure, the Higgs mechanism, and 't Hooft's proof of renormalizability, the Standard Model was complete: $SU(3)_c \times SU(2)_L \times U(1)_Y$ with spontaneous breaking to $SU(3)_c \times U(1)_{\text{em}}$, giving masses to $W^\pm$, Z^0 , and the fermions through Yukawa couplings to the Higgs field.

Yang and Mills' paper provided the mathematical foundation. Pauli's question identified the central physical obstacle. The Higgs mechanism and 't Hooft's renormalization removed it. The commutator term $-ig[B_\mu, B_\nu]$ that Yang and Mills added to the field strength—the term forced by the non-commutativity of $SU(2)$ —is the origin of gluon self-interaction, asymptotic freedom, confinement, and the entire structure of the strong and electroweak forces.

	$U(1)$ (QED)	$SU(2)$ (Yang–Mills)
Gauge field	A_μ (1 component)	B_μ^a (3 components)
Generator	1 (trivial)	$\tau^a = \sigma^a/2$ (non-commuting)
Commutator $[\tau^a, \tau^b]$	0	$i\epsilon^{abc}\tau^c$
Field strength	$\partial_\mu A_\nu - \partial_\nu A_\mu$	$\partial_\mu B_\nu - \partial_\nu B_\mu - ig[B_\mu, B_\nu]$
Self-interaction	No	Yes (3- and 4-boson vertices)
Gauge transformation	$A_\mu + \partial_\mu \chi$ (additive)	$UB_\mu U^\dagger + (i/g)(\partial_\mu U)U^\dagger$ (conjugation)
Mass term allowed?	No ($m^2 A_\mu A^\mu$ not invariant)	No ($m^2 B_\mu B^\mu$ not invariant)
Mass mechanism	Photon is massless	Higgs mechanism (1964)
Renormalizable?	Yes (trivially)	Yes ('t Hooft, 1971)

30.10 Connection and curvature: what the gauge field is

The gauge field A_μ (for $U(1)$) or B_μ^a (for $SU(2)$, $SU(3)$) is often introduced as “just another field” in the Lagrangian. But it has a geometric meaning that illuminates why it must exist, why it transforms the way it does, and why the field strength $F_{\mu\nu}$ has the form it has. We explain this using only the mathematics already in this book—no fibre bundles, no differential forms, no abstract geometry.

The problem: comparing phases at different points

Consider a charged field $\psi(x)$ with the local gauge symmetry $\psi(x) \rightarrow e^{i\alpha(x)}\psi(x)$. The phase of ψ at each point is unphysical: we can rotate it by a different amount at each point without changing any observable ($|\psi|^2$ is invariant).

But now consider the *derivative* $\partial_\mu\psi$. The derivative compares ψ at two nearby points x and $x + dx$:

$$\partial_\mu\psi(x) = \lim_{dx \rightarrow 0} \frac{\psi(x + dx) - \psi(x)}{dx^\mu}. \quad (30.58)$$

The problem: $\psi(x + dx)$ and $\psi(x)$ have been gauge-rotated by different phases ($\alpha(x + dx)$ and $\alpha(x)$). Subtracting them mixes two quantities defined with different phase conventions. The naive derivative $\partial_\mu\psi$ is *not* gauge-covariant:

$$\partial_\mu\psi \rightarrow \partial_\mu(e^{i\alpha}\psi) = e^{i\alpha}(\partial_\mu\psi + i(\partial_\mu\alpha)\psi) \neq e^{i\alpha}\partial_\mu\psi. \quad (30.59)$$

The extra term $i(\partial_\mu\alpha)\psi$ is the mismatch between the phase conventions at x and $x + dx$.

This is not a technicality. The kinetic term $|\partial_\mu\psi|^2$ in the Lagrangian is not gauge-invariant if $\partial_\mu\psi$ is not gauge-covariant. The theory is inconsistent unless we fix this.

The connection: a rule for transporting phase

The solution is to introduce a rule for “transporting” the phase of ψ from x to $x + dx$ before comparing. Define the *parallel transport* of $\psi(x)$ to the point $x + dx$ as:

$$\psi(x) \xrightarrow{\text{transport to } x+dx} e^{-ieA_\mu(x)dx^\mu}\psi(x) \approx (1 - ieA_\mu dx^\mu)\psi(x). \quad (30.60)$$

The factor $e^{-ieA_\mu dx^\mu}$ is a phase rotation that adjusts $\psi(x)$ to the phase convention at $x + dx$. The gauge field A_μ is the *connection*: it specifies, at each point and in each direction, how much the phase must be rotated to make the comparison meaningful.

The covariant derivative compares $\psi(x + dx)$ with the transported $\psi(x)$:

$$\begin{aligned} D_\mu\psi(x) &= \lim_{dx \rightarrow 0} \frac{\psi(x + dx) - e^{-ieA_\mu dx^\mu}\psi(x)}{dx^\mu} \\ &= \partial_\mu\psi(x) + ieA_\mu(x)\psi(x) \\ &= (\partial_\mu + ieA_\mu)\psi(x). \end{aligned} \quad (30.61)$$

The covariant derivative subtracts not $\psi(x)$ itself but the version of $\psi(x)$ that has been transported to $x + dx$ using the connection A_μ . The extra term $ieA_\mu\psi$ compensates for the phase mismatch, making $D_\mu\psi$ transform covariantly:

$$D_\mu\psi \rightarrow e^{i\alpha}D_\mu\psi. \quad (30.62)$$

No extra term, no mismatch. The kinetic term $|D_\mu\psi|^2$ is gauge-invariant.

The connection is what makes comparison possible. Without the connection A_μ , we cannot compare ψ at different points in a gauge-invariant way. The connection is the minimum additional structure needed to define a derivative that respects the local symmetry. In this sense, the gauge field is not an optional addition to the theory—it is *forced* by the requirement that the derivative makes sense.

Analogy: currency exchange rates. Imagine a world where each city has its own currency. The “field” $\psi(x)$ is a price, denominated in the local currency at city x . Comparing prices at two cities requires a currency exchange rate: the connection $A_\mu(x)$ in direction μ tells you how to convert prices at x to the currency at $x + dx$. Without the exchange rate, the raw price difference $\psi(x + dx) - \psi(x)$ is meaningless—you are subtracting dollars from euros. The covariant derivative $D_\mu\psi$ converts $\psi(x)$ to the currency at $x + dx$ before subtracting: a meaningful comparison.

A gauge transformation $\psi \rightarrow e^{i\alpha}\psi$ is a change of local currency: redenominating at each city. The exchange rate (connection) must adjust accordingly: $A_\mu \rightarrow A_\mu + \frac{1}{e}\partial_\mu\alpha$. The physical content—whether goods are cheaper in one city than another—is independent of the currency convention.

Curvature: what happens around a closed loop

The connection A_μ tells us how to transport phase along infinitesimal steps. Now ask: what happens when we transport phase around a *closed loop*?

Transport ψ from x around a small rectangle in the (μ, ν) plane with sides dx^μ and dx^ν :

1. From x to $x + dx^\mu \hat{e}_\mu$: phase factor $e^{-ieA_\mu(x)dx^\mu}$.
2. From there to $x + dx^\mu \hat{e}_\mu + dx^\nu \hat{e}_\nu$: phase factor $e^{-ieA_\nu(x+dx^\mu \hat{e}_\mu)dx^\nu}$.
3. Back along $-\mu$: phase factor $e^{+ieA_\mu(x+dx^\nu \hat{e}_\nu)dx^\mu}$.
4. Back along $-\nu$ to x : phase factor $e^{+ieA_\nu(x)dx^\nu}$.

The total phase acquired around the loop is the product of these four factors. Expanding to leading order in $dx^\mu dx^\nu$:

$$\text{total phase} \approx \exp(-ieF_{\mu\nu}(x) dx^\mu dx^\nu), \quad (30.63)$$

where:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \quad (30.64)$$

If $F_{\mu\nu} = 0$, the phase acquired around any closed loop is zero: transport is *path-independent*, and the connection is “flat.” If $F_{\mu\nu} \neq 0$, the phase depends on the path: the connection has *curvature*.

The field strength $F_{\mu\nu}$ is the *curvature* of the connection A_μ : it measures how much the parallel transport around a closed loop differs from the identity.

The electromagnetic field is curvature. The electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ are the components of $F_{\mu\nu}$: $F_{0i} = E_i$, $F_{ij} = -\epsilon_{ijk}B_k$. An electron moving in a magnetic field acquires a phase $e^{-ie \oint A_\mu dx^\mu} = e^{-ie \int F_{\mu\nu} dS^{\mu\nu}}$ (by Stokes’ theorem): the Aharonov–Bohm phase. The electromagnetic field is literally the curvature of the $U(1)$ connection A_μ .

Non-Abelian curvature: the commutator appears

For a non-Abelian gauge group ($SU(2)$, $SU(3)$), the parallel transport factor $e^{-igB_\mu dx^\mu}$ is a matrix, not a phase. Matrices do not commute: the order of the four transport steps around the rectangle matters.

Repeating the loop calculation with matrices:

$$\text{total transformation} \approx \exp(-igG_{\mu\nu} dx^\mu dx^\nu), \quad (30.65)$$

where:

$$G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu - ig[B_\mu, B_\nu]. \quad (30.66)$$

The commutator $[B_\mu, B_\nu]$ arises because the transport along μ then ν is not the same as along ν then μ : the matrices $e^{-igB_\mu dx^\mu}$ and $e^{-igB_\nu dx^\nu}$ do not commute, and their commutator contributes to the curvature at order $dx^\mu dx^\nu$.

For $U(1)$: $[A_\mu, A_\nu] = 0$ (numbers commute), and (30.66) reduces to $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. For $SU(3)_c$: $[B_\mu, B_\nu] \neq 0$ (matrices do not commute), and the curvature contains the extra commutator term that Yang and Mills discovered.

The Yang–Mills commutator term is curvature. The term $-ig[B_\mu, B_\nu]$ in the field strength (30.66) is not an ad hoc addition—it is the inevitable consequence of non-commutative parallel transport around a closed loop. The gluon self-interaction (three-gluon and four-gluon vertices) arises because the curvature of a non-Abelian connection is nonlinear in the connection.

Connection and curvature: the dictionary

Geometry	Gauge theory	Physics
Connection	Gauge field A_μ or B_μ^a	Photon, gluon, W/Z
Parallel transport	Phase factor $e^{-ieA_\mu dx^\mu}$	How phase changes along a path
Covariant derivative	$D_\mu = \partial_\mu - igB_\mu$	Derivative that respects gauge symmetry
Curvature	Field strength $F_{\mu\nu}$ or $G_{\mu\nu}^a$	E and B fields
Flat connection ($F = 0$)	Pure gauge ($A_\mu = \frac{1}{e}\partial_\mu\alpha$)	No electromagnetic field
Curvature $\neq 0$	$F_{\mu\nu} \neq 0$	Electromagnetic or chromodynamic field present
Holonomy (loop phase)	$e^{-ie\oint A\cdot dx}$	Aharonov–Bohm phase
Non-Abelian curvature	$G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu - ig[B_\mu, B_\nu]$	Gluon self-interaction

The analogy with general relativity

The same connection-and-curvature structure appears in general relativity, and the analogy is exact:

Concept	Gauge theory	General relativity
What is compared	Phase of ψ at different x	Direction of a vector at different x
Symmetry group	$U(1), SU(2), SU(3)$	$GL(4, \mathbb{R})$ (general coordinate)
Connection	A_μ, B_μ^a	Christoffel symbol $\Gamma_{\mu\nu}^\rho$
Covariant derivative	$D_\mu\psi = (\partial_\mu - igB_\mu)\psi$	$\nabla_\mu V^\rho = \partial_\mu V^\rho + \Gamma_{\mu\nu}^\rho V^\nu$
Curvature	$F_{\mu\nu}, G_{\mu\nu}^a$	Riemann tensor $R^\rho{}_{\sigma\mu\nu}$
Flat = no force	$F_{\mu\nu} = 0$: no EM field	$R^\rho{}_{\sigma\mu\nu} = 0$: flat spacetime
Equation of motion	Maxwell / Yang–Mills	Einstein: $G_{\mu\nu} = 8\pi GT_{\mu\nu}$

In general relativity, the Christoffel symbol $\Gamma_{\mu\nu}^\rho$ is the connection: it tells you how to transport a vector from x to $x + dx$ along curved spacetime. The Riemann tensor $R^\rho{}_{\sigma\mu\nu}$ is the curvature: it measures how much a vector rotates when transported around a closed loop. Einstein’s equation $G_{\mu\nu} = 8\pi GT_{\mu\nu}$ relates the curvature to the matter content, just as Maxwell’s equation $\partial^\mu F_{\mu\nu} = J_\nu$ relates the gauge curvature to the current.

The parallel is not a metaphor. It is a mathematical identity: both gauge theory and general relativity are theories of connections on spaces with local symmetry, and the field equations in both cases relate the curvature of the connection to the sources that create it. Weyl was the first to see this parallel clearly (in his 1918 and 1929 papers), and it remains the deepest structural insight connecting gauge theory to gravity.

Why curvature gives rise to force. In general relativity, a particle moving through curved spacetime follows a geodesic—the straightest possible path—but the curvature makes the geodesic deviate from a straight line: this deviation is gravity. In gauge theory, a charged particle moving through a region of nonzero $F_{\mu\nu}$ acquires a path-dependent phase, and the gradient of this phase produces the Lorentz force $f^\mu = eF^{\mu\nu}u_\nu$. In both cases, force is the physical manifestation of curvature: the connection tells the particle how to move; the curvature makes it deviate.

31 The Higgs Mechanism: Smuggling Mass with Scalar Mole

Mass is forbidden in gauge theories. This is not an accident or an approximation—it is a direct consequence of the symmetry structure. The Higgs mechanism is the discovery of a loophole: a spin-0 field with an imaginary mass can absorb the symmetry-breaking cost on behalf of the other fields, smuggling mass into a theory that would otherwise forbid it. Spontaneous symmetry breaking and the eating of Goldstone bosons are consequences of this manoeuvre, not the cause of it.

31.1 What is mass in quantum field theory?

Before asking how the Higgs mechanism generates mass, we need to be precise about what mass *is* in quantum field theory. The answer is not the familiar Newtonian “resistance to acceleration.” In QFT, mass has three equivalent faces: a parameter in the Lagrangian, the curvature of the potential, and the cost of creating an excitation at rest.

Face 1: mass as a Lagrangian parameter

In the Lagrangian for a free scalar field:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu\phi)^2 - \frac{1}{2}m^2\phi^2, \quad (31.1)$$

the parameter m appears as the coefficient of $-\frac{1}{2}\phi^2$. The Euler–Lagrange equation is the Klein–Gordon equation:

$$(\partial_\mu\partial^\mu + m^2)\phi = 0, \quad (31.2)$$

which in momentum space gives the dispersion relation:

$$k^2 = k_\mu k^\mu = E^2 - |\mathbf{p}|^2 = m^2. \quad (31.3)$$

The parameter m is the invariant mass: the energy of the particle at rest ($\mathbf{p} = 0$, $E = m$). For fermions the mass enters as $m\bar{\Psi}\Psi$; for gauge bosons as $\frac{1}{2}m^2 A_\mu A^\mu$; but the dispersion relation (31.3) is universal.

Massless fields ($m = 0$) satisfy $E = |\mathbf{p}|$ and propagate at the speed of light. Massive fields ($m > 0$) satisfy $E = \sqrt{|\mathbf{p}|^2 + m^2} > |\mathbf{p}|$ and can be brought to rest.

Face 2: mass as curvature of the potential

The potential $V(\phi) = \frac{1}{2}m^2\phi^2$ is a parabola centred at $\phi = 0$. The mass is the curvature at the minimum:

$$m^2 = \left. \frac{d^2V}{d\phi^2} \right|_{\phi=0}. \quad (31.4)$$

This is universal: for any potential $V(\phi)$ with a minimum at ϕ_0 , the mass of the physical fluctuation around that minimum is:

$$m_{\text{phys}}^2 = \left. \frac{d^2V}{d\phi^2} \right|_{\phi=\phi_0}. \quad (31.5)$$

The curvature controls how hard it is to perturb the field:

- **Sharp bowl** (large m^2): small displacement, large restoring force, rapid oscillation. Heavy particle, expensive to create.
- **Flat bowl** (small m^2): small restoring force, slow oscillation. Light particle, cheap to create.

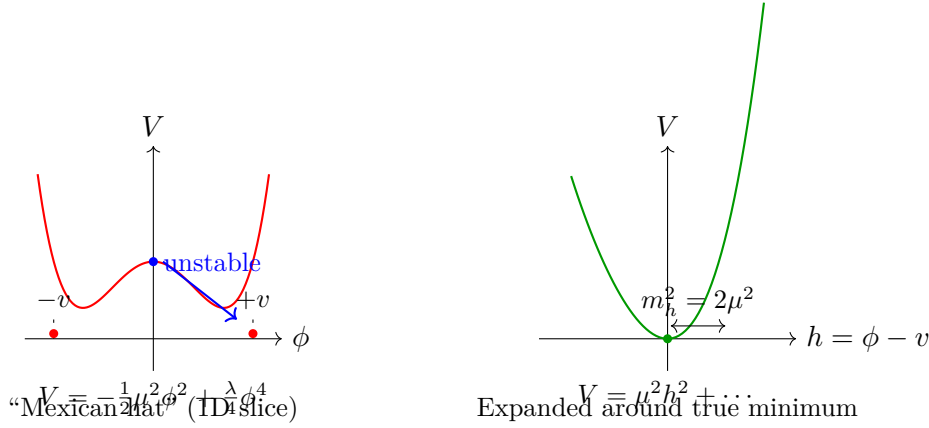
- **Flat direction** ($m^2 = 0$): no restoring force, zero energy cost. Massless particle—a Goldstone boson.
- **Inverted bowl** ($m^2 < 0$): anti-restoring force, the field rolls away. Imaginary mass—a tachyon—signals instability, not a physical particle.

Perturbing around the wrong point gives the wrong mass. The physical mass is always the curvature at the *true* minimum. For the Higgs potential $V(\phi) = -\frac{1}{2}\mu^2\phi^2 + \frac{\lambda}{4}\phi^4$:

$$\left. \frac{d^2V}{d\phi^2} \right|_{\phi=0} = -\mu^2 < 0 \quad (\text{tachyon at the origin: unstable, not physical}), \quad (31.6)$$

$$\left. \frac{d^2V}{d\phi^2} \right|_{\phi=v} = 2\mu^2 > 0 \quad (\text{physical Higgs mass } m_h = \sqrt{2}\mu \text{ at the minimum}). \quad (31.7)$$

Expanding around the wrong point gives an imaginary mass; expanding around the true minimum gives the real physical mass. This is the key step that most treatments gloss over.



Face 3: mass as the energy gap in Fock space

In the Fock hidden state, mass is the minimum energy to create one particle at rest. The free Hamiltonian is:

$$H_{\text{free}} = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \omega_{\mathbf{k}} a^\dagger(\mathbf{k}) a(\mathbf{k}), \quad \omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}. \quad (31.8)$$

Creating one particle at rest ($\mathbf{k} = 0$) costs energy $\omega_0 = m$: the mass is the energy gap between the vacuum $|0\rangle$ and the one-particle state $|\mathbf{k} = 0\rangle$. A massless field has $\omega_0 = 0$; a massive field has a nonzero gap.

The three faces are consistent: the Lagrangian parameter m determines the potential curvature, the curvature determines the oscillation frequency $\omega_0 = m$ at rest, and the oscillation frequency is the energy cost of creating one quantum.

Remark. [ML analogy: the Hessian as a mass matrix] In a neural network, the loss landscape $\mathcal{L}(\theta)$ plays the role of the potential $V(\phi)$. A local minimum θ^* is the vacuum. The Hessian $\nabla^2 \mathcal{L}|_{\theta^*}$ is the curvature matrix at the minimum. Its eigenvalues are the squared masses of the fluctuation modes: large eigenvalues are stiff directions (heavy particles, expensive to excite), small eigenvalues are flat directions (light particles), zero eigenvalues are exactly flat (massless Goldstone modes). Sharp minima generalise poorly (heavy modes, large loss curvature); flat minima generalise well (light modes, small loss curvature). The connection is exact: the Hessian of the loss at the minimum is the mass matrix of the physical fluctuations.

31.2 Why mass is forbidden for spin- $\frac{1}{2}$ and spin-1, but allowed for spin-0

The three particle species of the Standard Model — spin-0 scalars, spin- $\frac{1}{2}$ fermions, and spin-1 gauge bosons — behave completely differently with respect to mass terms. The contrast is worth stating all at once before explaining each case, because it is the contrast that makes the Higgs mechanism both necessary and possible.

Spin	Mass term	Gauge invariant?	Reason
0 (scalar)	$m^2 \phi^* \phi$	✓ Yes	No indices; singlet is trivially invariant
$\frac{1}{2}$ (fermion)	$m \bar{\psi}_L \psi_R$	× No	Index mismatch between chiralities
1 (vector)	$m^2 A_\mu A^\mu$	× No	Inhomogeneous gauge shift produces extra terms

The spin-0 case is the unique exception. This is not accidental: it is precisely why the Higgs field must be spin-0. A spin-0 field is the only species that can carry a mass term without violating gauge invariance, and therefore the only species whose vacuum expectation value can generate mass for the others without breaking the symmetry of the theory.

Spin-0: mass is allowed.

A real scalar field ϕ with no spinor index and no gauge index (a gauge singlet) has mass term $-\frac{1}{2}m^2\phi^2$. Under any gauge transformation, a singlet scalar is simply unchanged: $\phi \rightarrow \phi$. The mass term is therefore automatically gauge invariant, with no condition on m whatsoever.

For a complex scalar ϕ^i in a non-trivial gauge representation, the mass term must be the gauge-invariant combination $m^2 \phi^{i*} \phi^i = m^2 |\phi|^2$, which is invariant because the transformation matrix U on ϕ^i is cancelled by its conjugate ($U^\dagger$) on ϕ^{i*} . No index is left uncontracted; the mass term is allowed.

The key observation is that a scalar has *no Lorentz indices* and *no spinor structure*. The forbidden-mass arguments for fermions and gauge bosons both depend on specific index problems that a scalar simply does not have.

Spin- $\frac{1}{2}$: mass is forbidden by index mismatch.

A Dirac mass term couples left- and right-chiral fields:

$$\mathcal{L}_{\text{mass}}^{(1/2)} = m \bar{\Psi} \Psi = m (\bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L). \quad (31.9)$$

In the electroweak theory, ψ_L is an $SU(2)_L$ doublet with isospin index $i \in \{1, 2\}$, while ψ_R is an $SU(2)_L$ singlet with no isospin index. The product $\bar{\psi}_{L,i} \psi_R$ leaves a free isospin index i on $\bar{\psi}_L$ with no partner to contract against. Under $V \in SU(2)_L$:

$$\bar{\psi}_{L,i} \psi_R \mapsto \bar{\psi}_{L,j} (V^\dagger)^j_i \psi_R. \quad (31.10)$$

The factor $(V^\dagger)^j_i$ cannot be cancelled. The mass term transforms non-trivially and is forbidden.

The root cause is an *index mismatch*: ψ_L and ψ_R carry different $SU(2)_L$ representations (doublet and singlet respectively), so their isospin indices cannot be contracted. In QCD, by contrast, both chiralities carry the *same* $SU(3)_c$ representation, so $\bar{\psi}_{L,\alpha} \psi_R^\alpha$ contracts to a singlet and the quark mass term is allowed.

Spin-1: mass is forbidden by the inhomogeneous shift.

A mass term for a gauge boson A_μ would be $\frac{1}{2}m^2 A_\mu A^\mu$. Under $A_\mu \rightarrow A_\mu + \partial_\mu \alpha$:

$$A_\mu A^\mu \mapsto (A_\mu + \partial_\mu \alpha)(A^\mu + \partial^\mu \alpha) = A_\mu A^\mu + 2A_\mu \partial^\mu \alpha + (\partial_\mu \alpha)^2. \quad (31.11)$$

The extra terms involving $\partial_\mu \alpha$ do not vanish. For non-Abelian gauge bosons, the inhomogeneous shift $\frac{i}{g}(\partial_\mu U)U^\dagger$ in $G_\mu \rightarrow UG_\mu U^\dagger + \frac{i}{g}(\partial_\mu U)U^\dagger$ similarly prevents any quadratic term in G_μ from being gauge invariant.

The fundamental distinction between the fermion and gauge boson cases is the *type* of gauge transformation: matter fields transform homogeneously ($\psi \rightarrow U\psi$, so $U^\dagger U = I$ can cancel), while gauge fields transform inhomogeneously (with an additional derivative term that cannot be cancelled by any polynomial in the field).

The loophole, stated precisely. Because a spin-0 field is the only species that can carry a gauge-invariant mass term, it is the only species whose dynamics can be arranged to generate masses for the others. Specifically: if a spin-0 field acquires a nonzero vacuum expectation value, then:

- Its **Yukawa coupling** $y\phi\bar{\psi}_L\psi_R$ to fermions — gauge invariant because ϕ supplies the missing isospin index — becomes $yv\bar{\psi}_L\psi_R$ after the vacuum shift: a fermion mass.
- Its **kinetic term** $|D_\mu\phi|^2$ — gauge invariant by construction — produces $g^2v^2A_\mu A^\mu$ after the vacuum shift: a gauge boson mass.

The spin-0 field acts as the carrier in both cases: it absorbs the symmetry-breaking cost through its vacuum shift, and delivers a mass term that would otherwise be directly forbidden. The Higgs field is spin-0 not by convention but by necessity: no other spin could do this job.

31.3 Mass in QCD alone: quarks yes, gluons no

Before the full electroweak theory, it is illuminating to ask which mass terms are allowed by $SU(3)_c$ gauge invariance alone.

Quark mass: allowed

The quark mass term:

$$\mathcal{L}_{\text{quark mass}} = -m_q \bar{\psi}_\alpha \psi^\alpha = -m_q \sum_{\alpha \in \{r,g,b\}} \bar{\psi}_\alpha \psi^\alpha \quad (31.12)$$

contracts the colour index between $\bar{\psi}$ and ψ . Under $U \in SU(3)_c$:

$$\bar{\psi}_\alpha \psi^\alpha \mapsto \bar{\psi}_\beta (U^\dagger)^\beta_\alpha U^\alpha_{\beta'} \psi^{\beta'} = \bar{\psi}_\beta (U^\dagger U)^\beta_{\beta'} \psi^{\beta'} = \bar{\psi}_\alpha \psi^\alpha. \quad (31.13)$$

The mass term is exactly gauge invariant: $U^\dagger U = I$ absorbs both matrices. This works because both ψ_L^α and ψ_R^α carry the same $SU(3)_c$ representation (the fundamental **3**), so their colour indices match and contract.

The contrast with the electroweak case is stark. In QCD, both chiralities carry the same gauge representation, so the colour indices of $\bar{\psi}_L$ and ψ_R can contract to a singlet. In $SU(2)_L$, the chiralities carry *different* representations (doublet vs singlet), so the contraction fails.

Theory	ψ_L rep	ψ_R rep	Mass term
$SU(3)_c$	3	3	$\bar{\psi}_\alpha \psi^\alpha$: gauge-invariant ✓
$SU(2)_L$	2 doublet	1 singlet	$\bar{\psi}_L^i \psi_{Ri}$: not gauge-invariant ✗

Chiral symmetry of massless QCD. Setting $m_q = 0$ decouples ψ_L and ψ_R completely. The massless QCD Lagrangian is invariant under independent global $SU(N_f)$ rotations of N_f quark flavours in each chirality: $\psi_{Lf} \rightarrow (U_L)^f_{f'} \psi_{Lf'}$ and $\psi_{Rf} \rightarrow (U_R)^f_{f'} \psi_{Rf'}$ independently. This is

the global chiral symmetry $SU(N_f)_L \times SU(N_f)_R$. The quark mass term explicitly breaks it by coupling the two chiralities. For the light quarks (u, d, s) with masses much smaller than $\Lambda_{\text{QCD}} \approx 200 \text{ MeV}$, chiral symmetry is a good approximate symmetry of the strong interactions.

Gluon mass: still forbidden

Despite allowing quark mass, $SU(3)_c$ forbids gluon mass. The mass term $\frac{1}{2}m_g^2 G_\mu^a G^{a\mu}$ fails because the gluon field transforms *inhomogeneously*:

$$G_\mu \mapsto U G_\mu U^\dagger + \frac{i}{g_s}(\partial_\mu U)U^\dagger. \quad (31.14)$$

The inhomogeneous piece $\partial_\mu U$ generates cross terms in $G_\mu^a G^{a\mu}$ that do not cancel. The fundamental distinction:

Field	Transformation	Type	Mass term
Quark ψ^α	$U^\alpha_\beta \psi^\beta$	Homogeneous	Allowed
Gluon G_μ^a	$U G U^\dagger + \frac{i}{g} \partial U \cdot U^\dagger$	Inhomogeneous	Forbidden

Homogeneous transformations allow invariant bilinears; inhomogeneous ones do not. Gluons are therefore exactly massless in QCD. Despite this, quarks and gluons are confined and the strong force has a finite range—set not by a gluon mass but by the dynamically generated scale Λ_{QCD} , arising from the running coupling and gluon self-interaction.

31.4 The loophole: imaginary mass for a scalar mole

A real scalar field has no spinor index and no gauge index (if it is a singlet). Its mass term $-\frac{1}{2}m^2\phi^2$ is automatically gauge invariant with no obstruction. A scalar *can* have a mass, and can serve as the mole for smuggling.

Now exploit a loophole: nothing forbids a scalar mass with the wrong sign:

$$\mathcal{L}_{\text{Higgs}} = \frac{1}{2}(\partial_\mu \phi)^2 + \frac{1}{2}\mu^2\phi^2 - \frac{\lambda}{4}\phi^4, \quad \mu^2 > 0, \lambda > 0. \quad (31.15)$$

The sign flip $-\frac{1}{2}m^2\phi^2 \rightarrow +\frac{1}{2}\mu^2\phi^2$ corresponds to an imaginary mass $m = i\mu$. The potential $V(\phi) = -\frac{1}{2}\mu^2\phi^2 + \frac{\lambda}{4}\phi^4$ has the origin as a local *maximum*: the field cannot stay at $\phi = 0$.

The true minimum satisfies $V'(\phi) = 0$:

$$-\mu^2\phi + \lambda\phi^3 = 0 \quad \Rightarrow \quad \phi^2 = \frac{\mu^2}{\lambda} \equiv v^2. \quad (31.16)$$

The field settles at $|\phi| = v$, the *vacuum expectation value*.

Expanding around the true minimum. Define $h(x) = \phi(x) - v$, so $\langle h \rangle = 0$. Substituting into (31.15):

$$V(v+h) = \mu^2 h^2 + \sqrt{\lambda}\mu h^3 + \frac{\lambda}{4}h^4 + \text{const.} \quad (31.17)$$

The wrong-sign $+\frac{1}{2}\mu^2\phi^2$ has become a correct-sign $+\mu^2 h^2 = \frac{1}{2}(2\mu^2)h^2$: the Higgs mass is $m_h^2 = 2\mu^2$. The imaginary mass of ϕ at $\phi = 0$ has become the real mass of h at $\phi = v$. The loophole: *an imaginary mass forces a nonzero vacuum, and expanding around that vacuum reveals a real massive field.*

31.5 Goldstone's theorem and the fate of flat directions

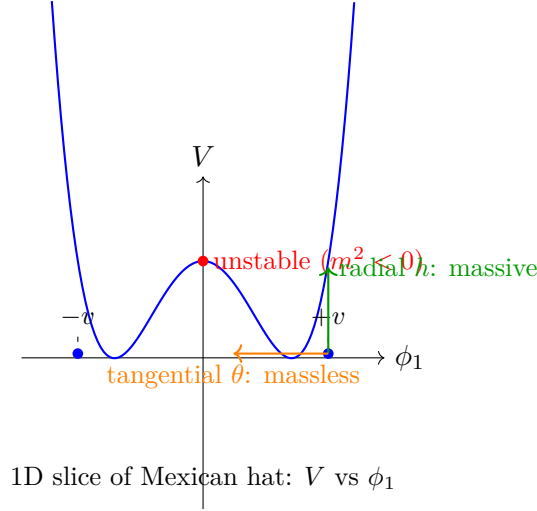
The shift $\phi = 0 \rightarrow \phi = v$ is a discrete choice for a real scalar. For a *complex* or multi-component scalar, the situation is richer and the concept of Goldstone bosons becomes essential.

Global symmetry and its breaking

Consider a complex scalar $\Phi = (\phi_1 + i\phi_2)/\sqrt{2}$ with a global $U(1)$ symmetry $\Phi \rightarrow e^{i\alpha}\Phi$ and potential:

$$V(\Phi) = -\mu^2|\Phi|^2 + \lambda|\Phi|^4. \quad (31.18)$$

This is the Mexican hat potential (Figure ??). The minima form a circle $|\Phi|^2 = \mu^2/2\lambda \equiv v^2/2$: there is a whole ring of degenerate vacua, parametrised by the phase $\theta \in [0, 2\pi)$.



The vacuum picks one point on the ring—say $\Phi_0 = v/\sqrt{2}$ (real). Write the fluctuations around this point in polar form:

$$\Phi(x) = \frac{1}{\sqrt{2}}(v + h(x)) e^{i\theta(x)/v}. \quad (31.19)$$

Here $h(x)$ is the *radial* fluctuation (away from the ring of minima) and $\theta(x)$ is the *angular* fluctuation (along the ring). Substituting into (31.18):

- The radial field h has potential $V \supset \mu^2 h^2$: positive curvature, mass $m_h^2 = 2\mu^2 > 0$. It costs energy to push Φ away from the ring. **h is massive.**
- The angular field θ has no potential term at all: the ring of minima is exactly flat, and sliding along it costs zero energy. **θ is massless.**

The massless θ is the *Goldstone boson*.

Goldstone's theorem

Theorem 31.1 (Goldstone, 1961). *For every continuous global symmetry that is spontaneously broken by the vacuum, there is one massless scalar particle—a Goldstone boson.*

The precise statement in terms of the potential: if G is a continuous symmetry group of the Lagrangian and the vacuum ϕ_0 is invariant only under a subgroup $H \subset G$, then there are $\dim G - \dim H$ massless scalars, one for each broken generator.

Why broken generators produce flat directions. A symmetry generator T that is broken by the vacuum satisfies $T\phi_0 \neq \phi_0$: it moves the vacuum to a different point in field space. But since G is a symmetry of the Lagrangian, $T\phi_0$ has the same energy as ϕ_0 : it is another minimum. Moving from ϕ_0 to $T\phi_0$ costs no energy. The direction $T\phi_0 - \phi_0$ in field space is therefore exactly flat—a zero-curvature direction—and the corresponding quantum is massless.

In our example: $G = U(1)$, $H = \{e\}$ (trivial, since any nonzero v breaks the $U(1)$), $\dim G - \dim H = 1 - 0 = 1$ Goldstone boson. This is the angular field θ .

The count for the Standard Model Higgs. The Higgs doublet ϕ^i ($i \in \{1, 2\}$, complex) has 4 real degrees of freedom. The symmetry group $G = SU(2)_L \times U(1)_Y$ has $\dim G = 3 + 1 = 4$ generators. The vacuum $\langle \phi \rangle = (0, v/\sqrt{2})^T$ is invariant under the $U(1)_{\text{em}}$ subgroup generated by $Q = T^3 + Y$ (the electric charge), which has $\dim H = 1$. The number of broken generators is $4 - 1 = 3$. By Goldstone's theorem: **3 Goldstone bosons.**

These three Goldstone bosons are exactly the three angular directions in the 4-dimensional field space of ϕ^i around the vacuum. They correspond to motions along the orbits of the three broken generators T^1 , T^2 , and $T^3 - Y$.

31.6 The Yukawa coupling for fermions: index structure

We now show explicitly how the isospin index of the Higgs doublet provides the missing partner that makes the Yukawa coupling gauge invariant — the partner whose absence made the direct mass term forbidden.

The coupling and its index structure. The Yukawa coupling for the electron is:

$$\mathcal{L}_{\text{Yukawa}} = -y_e \bar{L}_{L,i} H^i e_R + \text{h.c.}, \quad (31.20)$$

where $L_L^i = (\nu_{eL}, e_L)^T$ is the left lepton doublet with isospin index $i \in \{1, 2\}$, $H^i = (\phi^+, \phi^0)^T$ is the Higgs doublet, and e_R is the right-handed electron singlet (no isospin index). The notation $\bar{L}_{L,i}$ denotes the conjugate with the isospin index lowered: $\bar{L}_{L,i} = (\bar{L}_L^i)^\dagger \gamma^0$ with the index lowered by the identity in isospin space. The isospin index i is contracted between $\bar{L}_{L,i}$ (lower) and H^i (upper).

Under $V \in SU(2)_L$, the three fields transform as:

$$L_L^i \mapsto V^i_j L_L^j, \quad H^i \mapsto V^i_j H^j, \quad e_R \mapsto e_R \text{ (singlet)}. \quad (31.21)$$

Both L_L^i and H^i transform in the same (fundamental) doublet representation, so both acquire a factor of V^i_j . Taking the conjugate of $L_L^i \rightarrow V^i_j L_L^j$ gives $\bar{L}_{L,i} \rightarrow \bar{L}_{L,j} (V^\dagger)^j_i$. The Yukawa coupling then transforms as:

$$\bar{L}_{L,i} H^i \mapsto \bar{L}_{L,j} (V^\dagger)^j_i V^i_k H^k = \bar{L}_{L,j} (V^\dagger V)^j_k H^k = \bar{L}_{L,j} \delta_k^j H^k = \bar{L}_{L,i} H^i. \quad (31.22)$$

The product is gauge invariant: $(V^\dagger V)^j_k = \delta_k^j$ cancels the two transformation matrices exactly, because both fields carry the same doublet representation with matching upper and lower indices.

Why the direct mass term fails by the same logic. The forbidden fermion mass term $m \bar{L}_{L,i} e_R$ has exactly the same index structure on $\bar{L}_{L,i}$ — a lower isospin index. But e_R is a singlet: it provides no upper isospin index to contract against. Under $V \in SU(2)_L$:

$$\bar{L}_{L,i} e_R \mapsto \bar{L}_{L,j} (V^\dagger)^j_i e_R. \quad (31.23)$$

The factor $(V^\dagger)^j_i$ is present on $\bar{L}_{L,j}$ but there is nothing on the right to supply the compensating V^i_k . The free index i transforms non-trivially; the term is not gauge invariant. The Higgs doublet H^i is precisely the object that supplies the missing upper index i with the correct transformation law to cancel $(V^\dagger)^j_i$.

After electroweak symmetry breaking. In unitary gauge, the Higgs doublet takes the vacuum form:

$$H^i = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \quad \text{so} \quad \langle H^i \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \quad (31.24)$$

Substituting into (31.35) and evaluating the index contraction $\bar{L}_{L,i}\langle H^i\rangle$:

$$\bar{L}_{L,i}\langle H^i\rangle = \bar{L}_{L,1}\cdot 0 + \bar{L}_{L,2}\cdot \frac{v}{\sqrt{2}} = \frac{v}{\sqrt{2}}\bar{e}_L, \quad (31.25)$$

since $L_L^2 = e_L$ (the lower component of the doublet). The Yukawa coupling becomes:

$$-y_e \bar{L}_{L,i} H^i e_R|_{H=\langle H\rangle} = -\frac{y_e v}{\sqrt{2}} \bar{e}_L e_R + \text{h.c.} = -m_e \bar{e}_L e_R + \text{h.c.}, \quad m_e = \frac{y_e v}{\sqrt{2}}. \quad (31.26)$$

The neutrino $\nu_L = L_L^1$ picks up a zero from $\langle H^1\rangle = 0$ and remains massless at this order. The electron has acquired a mass $m_e = y_e v/\sqrt{2}$, entirely determined by the Yukawa coupling y_e (a free parameter of the theory) and the vacuum expectation value v (fixed by the Higgs potential parameters μ and λ).

For quarks, the same mechanism applies but with one additional subtlety: the up-type quarks (u , c , t) require the conjugate doublet $\tilde{H}^i = \varepsilon^{ij} H_j^*$ to form a gauge-invariant Yukawa coupling, since their hypercharge assignments require contracting with the complex-conjugate representation. The down-type quarks (d , s , b) use H^i directly, as in (31.35). In both cases, the pattern is identical: the Higgs doublet supplies the isospin index that would otherwise be uncontracted, and the vacuum expectation value converts the Yukawa coupling into a mass term.

The mass hierarchy. Every fermion mass in the Standard Model takes the form:

$$m_f = \frac{y_f v}{\sqrt{2}}, \quad (31.27)$$

with the same $v \approx 246$ GeV for all species. The observed mass hierarchy — from $m_e \approx 0.511$ MeV to $m_t \approx 173$ GeV — is entirely encoded in the spread of Yukawa couplings y_f , which range from $y_e \approx 2 \times 10^{-6}$ to $y_t \approx 1$. Why the Yukawa couplings have such a wide spread is one of the open questions of the Standard Model (the *flavour puzzle*); the Higgs mechanism explains the structure of the mass terms but does not predict the values of the couplings.

What happens with a local symmetry: the Goldstone bosons are eaten

For a global symmetry, the Goldstone boson θ in (31.19) is a physical massless particle. For a local symmetry $\Phi \rightarrow e^{i\alpha(x)}\Phi$, there is a gauge field A_μ that also transforms: $A_\mu \rightarrow A_\mu + \frac{1}{e}\partial_\mu\alpha$.

Under this local transformation, the angular field θ in (31.19) transforms as:

$$\frac{\theta(x)}{v} \mapsto \frac{\theta(x)}{v} + \alpha(x). \quad (31.28)$$

The Goldstone field $\theta(x)$ shifts by an arbitrary function of x —the same freedom as a gauge transformation. We can therefore choose $\alpha(x) = -\theta(x)/v$ to set $\theta(x) = 0$ everywhere. This is the *unitary gauge*: it uses the gauge freedom to remove the Goldstone boson from the field content entirely.

Where did the Goldstone boson go? It became the *longitudinal polarisation* of the gauge boson. To see this, substitute (31.19) (with $\theta = 0$ in unitary gauge) into the kinetic term $|D_\mu\Phi|^2$:

$$|D_\mu\Phi|^2|_{\theta=0} = \frac{1}{2}(\partial_\mu h)^2 + \frac{1}{2}e^2(v+h)^2 A_\mu A^\mu. \quad (31.29)$$

At leading order in h : the $e^2 v^2 A_\mu A^\mu$ term is a mass term for A_μ with $m_A = ev$. Before gauge fixing, the massless gauge boson A_μ had 2 physical polarisations (transverse). After gauge fixing, it has 3 polarisations (2 transverse + 1 longitudinal): the longitudinal mode is the eaten Goldstone boson.

The degree-of-freedom count is exact:

Before SSB	Degrees of freedom	After SSB
Massless A_μ	2 (transverse)	
Goldstone boson θ	1 (massless scalar)	
Radial mode h	1 (massive scalar)	
Total: 4		Total: 4
		Massive A_μ : 3 (2 transverse + 1 longitudinal)
		Physical Higgs h : 1 (massive scalar)

The Goldstone boson is not destroyed—it is *transmuted* into the longitudinal polarisation of the gauge boson. No degree of freedom is lost; the bookkeeping simply rearranges.

The eating is not the mechanism. The eating of the Goldstone boson is a consequence of the gauge symmetry and the nonzero vacuum, not the cause of the mass. The gauge boson acquires mass from the kinetic term (31.29) regardless of which gauge is chosen. The unitary gauge simply makes the physical content explicit by removing the redundant Goldstone degree of freedom. The real mechanism has three steps, in order:

1. A scalar is given an imaginary mass (wrong-sign potential).
2. This forces $\langle\phi\rangle = v \neq 0$.
3. The nonzero v , through Yukawa couplings and the scalar kinetic term, generates mass terms for fermions and gauge bosons.

The Goldstone bosons and their eating are bookkeeping for step 2 (broken continuous symmetry generates flat directions) and step 3 (local symmetry allows a gauge choice that removes the flat directions from the physical spectrum).

31.7 Smuggling mass into spin- $\frac{1}{2}$ fields: the Yukawa coupling

The nonzero vacuum $\langle\phi\rangle = v$ generates fermion masses through the Yukawa coupling:

$$\mathcal{L}_{\text{Yukawa}} = -y \phi \bar{\psi}_L \psi_R + \text{h.c.} \quad (31.30)$$

For a real scalar singlet ϕ , this is gauge invariant: ϕ has no indices, and the product $\bar{\psi}_L \psi_R$ can be arranged to be a singlet. Substituting $\phi = v + h$:

$$-y \phi \bar{\psi}_L \psi_R = \underbrace{-yv \bar{\psi}_L \psi_R}_{\text{Dirac mass } m_f=yv} + \underbrace{-yh \bar{\psi}_L \psi_R}_{\text{Higgs-fermion coupling}}. \quad (31.31)$$

The fermion has acquired a mass $m_f = yv$, proportional to both the Yukawa coupling and the vacuum expectation value.

The mechanism of the smuggle. The Dirac mass $m \bar{\psi}_L \psi_R$ is forbidden because it leaves an uncontracted gauge index. The Yukawa coupling $y \phi \bar{\psi}_L \psi_R$ is allowed because ϕ carries no gauge index—no index is left uncontracted. When ϕ shifts to its vacuum value, the Yukawa coupling becomes $yv \bar{\psi}_L \psi_R$, which *looks* like a mass term. The scalar has acted as a carrier: it absorbed the symmetry-breaking cost and delivered a fermion mass. The imaginary mass of ϕ is the engine; without it, $\langle\phi\rangle = 0$ and no fermion mass is generated.

31.8 Smuggling mass into spin-1 fields: the scalar kinetic term

For a complex scalar ϕ charged under $U(1)$, the gauge-invariant kinetic term is $|D_\mu \phi|^2 = |(\partial_\mu - igA_\mu)\phi|^2$. Substituting $\phi = v + h$ and keeping the leading term:

$$|D_\mu \phi|^2|_{\phi=v} \supset g^2 v^2 A_\mu A^\mu. \quad (31.32)$$

This is a mass term for A_μ with $m_A = gv$. It appeared from the kinetic term of the scalar—not from a forbidden mass term—because the scalar shifted to its vacuum value.

The mass term $\frac{1}{2}m^2 A_\mu A^\mu$ is forbidden. But $|D_\mu \phi|^2$ is gauge invariant by construction. When $\langle |\phi|^2 \rangle = v^2$, the kinetic term produces $g^2 v^2 A_\mu A^\mu$. The scalar's vacuum expectation value smuggled a mass into the gauge boson's kinetic term.

31.9 The full electroweak Higgs: a complex $SU(2)_L$ doublet

The Standard Model requires the scalar to be a complex $SU(2)_L$ doublet ϕ^i ($i \in \{1, 2\}$) with hypercharge $Y = +\frac{1}{2}$. Three requirements force this choice:

1. **Give mass to $W^\pm$ and Z^0 :** The covariant derivative must act non-trivially on ϕ , so ϕ must transform under $SU(2)_L$. A doublet ($i \in \{1, 2\}$) is the minimal choice.
2. **Give mass to fermions via Yukawa:** The isospin index of ϕ^i must contract with the isospin index of the left-handed fermion doublet, making the coupling gauge invariant.
3. **Leave the photon massless:** The hypercharge assignment must be chosen so that one linear combination of generators is unbroken by the vacuum.

The Lagrangian is:

$$\mathcal{L}_{\text{Higgs}} = |D_\mu \phi|^2 - V(\phi), \quad V(\phi) = -\mu^2 |\phi|^2 + \lambda |\phi|^4, \quad (31.33)$$

with covariant derivative:

$$D_\mu \phi^i = \partial_\mu \phi^i - ig W_\mu^a \left(\frac{\tau^a}{2}\right)^i_j \phi^j - ig' Y B_\mu \phi^i, \quad (31.34)$$

where $\tau^a = \sigma^a$ are the Pauli matrices (generators of $SU(2)_L$), g is the $SU(2)_L$ coupling, g' is the $U(1)_Y$ coupling, and $Y = +\frac{1}{2}$ is the hypercharge of ϕ .

The Yukawa coupling for fermions: index structure

Consider the Yukawa coupling for the electron:

$$\mathcal{L}_{\text{Yukawa}} = -y_e \bar{L}_L^i \phi_i e_R + \text{h.c.}, \quad (31.35)$$

where $L_L^i = (\nu_{eL}, e_L)^T$ is the left lepton doublet and e_R is the right-handed electron singlet. The isospin index i on $\bar{L}_L^i$ is contracted with the isospin index on $\phi_i = \phi^{i*}$ (the conjugate Higgs): one index up, one down, contraction to a singlet. The right-handed electron e_R has no isospin index.

Why is this gauge invariant? Under $V \in SU(2)_L$:

$$\bar{L}_L^i \phi_i \mapsto \bar{L}_{L,i'} (V^\dagger)^{i'}_i (V)^j_i \phi_j \cdot \text{(from } \phi_i \rightarrow V^j_i \phi_j \text{ and } \bar{L}_L^i \rightarrow \bar{L}_{L,i'} (V^\dagger)^{i'}_i) \quad (31.36)$$

Wait—let us be careful. Under $SU(2)_L$:

$$L_L^i \mapsto V^i_j L_L^j, \quad \phi^i \mapsto V^i_j \phi^j, \quad e_R \mapsto e_R \text{ (singlet, unchanged)}. \quad (31.37)$$

The Yukawa coupling using the conjugate doublet $\tilde{\phi}^i = \varepsilon^{ij} \phi_j^*$ (which also transforms as a doublet under $SU(2)_L$ due to the properties of $SU(2)$) allows us to form the invariant:

$$\bar{L}_L^i \tilde{\phi}_i e_R \mapsto \bar{L}_{L,j} (V^\dagger)^j_i (V)^i_k \tilde{\phi}^k e_R = \bar{L}_{L,j} \delta^j_k \tilde{\phi}^k e_R = \bar{L}_L^j \tilde{\phi}_j e_R. \quad (31.38)$$

The $(V^\dagger)^j_i V^i_k = (V^\dagger V)^j_k = \delta^j_k$ makes the product gauge invariant. Compare this to the forbidden mass term $m \bar{L}_L^i e_R$: the isospin index i on $\bar{L}_L^i$ has nothing to contract against, so it transforms nontrivially and the term is not gauge invariant. The Higgs doublet provides the missing isospin partner.

After the Higgs acquires its vacuum expectation value in unitary gauge:

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad (31.39)$$

the Yukawa coupling (31.35) becomes:

$$-y_e \bar{L}_L^i \langle \phi \rangle_i e_R = -\frac{y_e v}{\sqrt{2}} \bar{e}_L e_R + \text{h.c.} = -m_e \bar{e}_L e_R + \text{h.c.}, \quad m_e = \frac{y_e v}{\sqrt{2}}. \quad (31.40)$$

The electron mass is proportional to the Yukawa coupling y_e and the vacuum expectation value v . Heavier fermions have larger Yukawa couplings: the top quark, with $m_t \approx 173 \text{ GeV}$, has $y_t \approx 1$ (the largest of all Standard Model Yukawa couplings).

The gauge boson masses: explicit derivation

In unitary gauge, $\phi = \frac{1}{\sqrt{2}}(0, v+h)^T$. Substituting into the kinetic term $|D_\mu \phi|^2$ and keeping only the terms quadratic in gauge fields (setting $h=0$ for the mass terms):

$$\begin{aligned} |D_\mu \phi|^2|_{h=0} &= \left| \left(-ig \frac{\tau^a}{2} W_\mu^a - ig' \frac{1}{2} B_\mu \right) \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \right|^2 \\ &= \frac{v^2}{8} \left| \begin{pmatrix} gW_\mu^3 + g'B_\mu & g(W_\mu^1 - iW_\mu^2) \\ g(W_\mu^1 + iW_\mu^2) & -gW_\mu^3 + g'B_\mu \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right|^2. \end{aligned} \quad (31.41)$$

Evaluating the matrix-vector product:

$$|D_\mu \phi|^2|_{h=0} = \frac{g^2 v^2}{4} |W_\mu^+|^2 + \frac{g^2 v^2}{4} |W_\mu^-|^2 + \frac{v^2}{8} (gW_\mu^3 - g'B_\mu)^2, \quad (31.42)$$

where $W_\mu^\pm = (W_\mu^1 \mp iW_\mu^2)/\sqrt{2}$. Reading off the mass terms:

- **$W^\pm$ mass:** $\mathcal{L} \supset \frac{1}{2} m_W^2 W_\mu^+ W_\mu^-$ gives $m_W = \frac{gv}{2}$.
- **Z^0 mass:** The combination $gW_\mu^3 - g'B_\mu$ appears in (31.42). normalizing:

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2 + g'^2}}, \quad m_Z = \frac{v}{2} \sqrt{g^2 + g'^2} = \frac{m_W}{\cos \theta_W}, \quad (31.43)$$

where the Weinberg angle is $\cos \theta_W = g/\sqrt{g^2 + g'^2}$.

- **Photon mass:** The orthogonal combination $A_\mu = (g'W_\mu^3 + gB_\mu)/\sqrt{g^2 + g'^2}$ does *not* appear in (31.42): it is absent, so $m_\gamma = 0$. The photon remains massless.

The Weinberg mixing between W_μ^3 and B_μ to give the physical Z^0 and γ is:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \sin \theta_W & \cos \theta_W \\ \cos \theta_W & -\sin \theta_W \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix}. \quad (31.44)$$

The ratio of masses:

$$\frac{m_W}{m_Z} = \cos \theta_W \approx 0.88, \quad (31.45)$$

is a firm prediction of the Higgs mechanism with a single doublet. Experimentally: $m_W/m_Z = 80.4/91.2 \approx 0.881$. The agreement is a direct test of the doublet structure.

Why the photon is massless: unbroken generator. The electric charge generator $Q = T^3 + Y$ leaves the vacuum invariant:

$$e^{i\alpha Q} \langle \phi \rangle = e^{i\alpha(T^3+Y)} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} = e^{i\alpha((-\frac{1}{2})+\frac{1}{2})} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} = \langle \phi \rangle. \quad (31.46)$$

The lower component of ϕ^i has $T^3 = -\frac{1}{2}$ and $Y = +\frac{1}{2}$, giving $Q = -\frac{1}{2} + \frac{1}{2} = 0$: it is electrically neutral. The vacuum is therefore neutral and invariant under $U(1)_{\text{em}}$. This unbroken generator corresponds to the massless photon.

The three broken generators T^1 , T^2 , $T^3 - Y$ each produce one Goldstone boson—which is eaten to become the longitudinal polarisation of W^+ , W^- , Z^0 respectively.

Generator	Broken?	Goldstone	Eaten by	Gives mass to
T^1	Yes	π^1	W^+ longitudinal	$m_W = gv/2$
T^2	Yes	π^2	W^- longitudinal	$m_W = gv/2$
$T^3 - Y$	Yes	π^3	Z^0 longitudinal	$m_Z = \sqrt{g^2 + g'^2} v/2$
$T^3 + Y = Q$	No	—	—	$m_\gamma = 0$

31.10 The reparametrisation: unitary gauge and Goldstone eating

The complex doublet ϕ^i has 4 real degrees of freedom. We can parametrise them as:

$$\phi(x) = e^{i\pi^a(x)\tau^a/v} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \quad (31.47)$$

where $\pi^a(x)$ ($a = 1, 2, 3$) are the three Goldstone bosons (angular fluctuations along the broken-generator directions) and $h(x)$ is the physical Higgs field (radial fluctuation).

Under an $SU(2)_L$ gauge transformation $V(x) = e^{i\alpha^a(x)\tau^a/2}$, the exponential prefactor in (31.47) transforms as:

$$e^{i\pi^a\tau^a/v} \mapsto V(x) e^{i\pi^a\tau^a/v}. \quad (31.48)$$

We can choose $V(x) = e^{-i\pi^a(x)\tau^a/v}$ to set the exponential to the identity everywhere. This is the *unitary gauge*:

$$\phi(x)|_{\text{unitary}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \quad (31.49)$$

The three Goldstone fields π^a have been gauge-transformed to zero. They have not disappeared—they have been *transmuted* into the longitudinal polarisations of $W^\pm$ and Z^0 , as shown in the degree-of-freedom table above.

The physical content of the unitary gauge. In the unitary gauge, all unphysical degrees of freedom have been eliminated. The remaining physical spectrum is:

- One real scalar $h(x)$: the physical Higgs boson.
- Three massive vector bosons $W^\pm$, Z^0 (each with 3 polarisations: 2 transverse + 1 longitudinal).
- One massless vector boson γ (2 transverse polarisations).
- Massive fermions (from Yukawa couplings).

The unitary gauge is the clearest gauge for identifying physical particles. It is less convenient for Feynman diagram calculations (the W and Z propagators are complicated by the $k_\mu k_\nu / m_V^2$ term from the longitudinal polarisation), but it makes the particle content transparent.

31.11 The Fock vacuum and the mass gap

The classical potential minimum and the quantum Fock vacuum are the same object in two languages.

Classical (Lagrangian)	Quantum (Fock space)
Minimum of $V(\phi)$ at $\phi_0 = v$	Ground state $ 0\rangle$ of H
Curvature $V''(v) = m_h^2$	Mass gap m_h above $ 0\rangle$
Flat direction ($V'' = 0$)	Goldstone boson (massless)
Negative curvature ($V'' < 0$)	Tachyon (unstable, not a true min.)
Fluctuation $h = \phi - v$	Higgs boson excitation $a_h^\dagger 0\rangle$
Vacuum energy $V(v)$	Zero-point energy (normal-ordered away)

After expanding around the true vacuum in unitary gauge, the free Hamiltonian is:

$$H_{\text{free}} = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \left[\omega_{\mathbf{k}}^h a_h^\dagger a_h + \omega_{\mathbf{k}}^W \sum_{\lambda=0,\pm 1} a_W^{\dagger\lambda} a_W^\lambda + \omega_{\mathbf{k}}^Z \sum_{\lambda=0,\pm 1} a_Z^{\dagger\lambda} a_Z^\lambda + |\mathbf{k}| \sum_{\lambda=\pm 1} a_\gamma^{\dagger\lambda} a_\gamma^\lambda + \sum_f \omega_{\mathbf{k}}^f (b_f^\dagger b_f + d_f^\dagger d_f) \right], \quad (31.50)$$

where all frequencies $\omega_{\mathbf{k}}^h = \sqrt{|\mathbf{k}|^2 + m_h^2}$, $\omega_{\mathbf{k}}^W = \sqrt{|\mathbf{k}|^2 + m_W^2}$, etc. are real and positive. The Hamiltonian is bounded below; the Schrödinger RNN is stable. This stability is the ultimate payoff of the Higgs mechanism.

The minimum energy to create any particle at rest:

Particle	Mass	Min. energy at rest
Photon γ	0	$ \mathbf{k} \rightarrow 0$ (any)
Electron e^-	0.511 MeV	0.511 MeV
Proton p	938 MeV	938 MeV
$W^\pm$	80.4 GeV	80.4 GeV
Z^0	91.2 GeV	91.2 GeV
Higgs h	125.1 GeV	125.1 GeV

At room temperature ($k_B T \approx 25$ meV), only photons and the lightest fermions are thermally accessible; every massive sector sits in its Fock vacuum. The story told by the Schrödinger RNN at everyday energies is atomic and optical physics. The W , Z , Higgs, and heavy quark sectors are frozen in $|0\rangle$, waiting for energies of tens of GeV to populate their modes.

Remark. [Note on the QCD mass gap] The photon and gluon are classically massless, so the theory has no strict mass gap in the perturbative sense. However, the physical spectrum of QCD (the hadronic spectrum) *does* have a mass gap: the lightest hadron is the pion with $m_\pi \approx 140$ MeV. This gap is non-perturbative in origin (it arises from confinement and chiral symmetry breaking, not from the Higgs mechanism) and is unrelated to the perturbative gluon mass. The statement that the lightest massive Standard Model particle is the electron refers to the perturbative particle spectrum, not the hadronic spectrum.

Summary of the Higgs mechanism.

Field	Why mass is forbidden	How mass is generated
Spin- $\frac{1}{2}$ fermion	$m\bar{\psi}_L\psi_R$ breaks $SU(2)_L$ (index mismatch)	Yukawa $y_e\bar{L}_L^i\phi_ie_R$: $m_f = y_f v/\sqrt{2}$
$W^\pm, Z^0$ bosons	$m^2 A_\mu A^\mu$ breaks gauge invariance	$ D_\mu\phi ^2$ at vacuum: $m_W = gv/2$, $m_Z = m_W/\cos\theta_W$
Photon γ	— (unbroken $Q = T^3 + Y$)	Remains massless: $m_\gamma = 0$
Gluons	Inhomogeneous $SU(3)_c$ transform	Not affected by Higgs: $m_g = 0$
Higgs h	No obstruction for scalar	Curvature at true minimum: $m_h = \sqrt{2}\mu = \sqrt{2\lambda}v$
<i>Engine of all fermion and weak boson masses: imaginary mass $\Rightarrow \langle\phi\rangle = v \neq 0 \Rightarrow$ masses.</i>		

32 Fock Space for Gauge Fields

In the scalar field section, we built the Fock space by starting with a single-particle Hilbert space spanned by momentum eigenstates $|\mathbf{k}\rangle$, then taking the direct sum over all particle numbers. The result was a space of states labelled by how many scalar particles are present and what their momenta are. The Hamiltonian then rotated the hidden state within this space.

For the Standard Model, the same construction applies, but each particle now carries additional quantum numbers: spin, colour, isospin, flavour. The Fock space grows by a direct product for each new index. This section builds that space systematically, from the single-particle state outward.

32.1 Enriching the single-particle state: direct products

Recap: scalar Fock space. For a free scalar field, the single-particle Hilbert space is spanned by momentum eigenstates:

$$|\mathbf{k}\rangle, \quad \mathbf{k} \in \mathbb{R}^3, \quad (32.1)$$

with inner product $\langle \mathbf{k} | \mathbf{k}' \rangle = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}')$. The n -particle space is the symmetrised (bosons) or antisymmetrised (fermions) tensor product of n copies of this single-particle space, and the full Fock space is the direct sum over all n :

$$\mathcal{F}_{\text{scalar}} = \mathbb{C} \oplus \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)} \oplus \dots \quad (32.2)$$

Adding a spin index. A Weyl spinor field carries a spinor index $s \in \{1, 2\}$ (from the $(\frac{1}{2}, 0)$ representation of $SL(2, \mathbb{C})$). Each single-particle state now carries both a momentum $\mathbf{k}$ and a spin label s :

$$|\mathbf{k}\rangle \mapsto |\mathbf{k}, s\rangle, \quad s \in \{1, 2\}. \quad (32.3)$$

The single-particle Hilbert space becomes:

$$\mathcal{H}_{\text{spin}}^{(1)} = \mathcal{H}_{\text{scalar}}^{(1)} \otimes \mathcal{H}_{\text{spin}}, \quad (32.4)$$

where $\mathcal{H}_{\text{spin}} \cong \mathbb{C}^2$ is the 2-dimensional spin space. This is a *direct product*: the new basis is all pairs $(\mathbf{k}, s)$, so the number of basis vectors doubles. Nothing changes about the direct sum structure over particle numbers; only the label on each particle is richer.

Adding a colour index. A quark field additionally carries a colour index $\alpha \in \{r, g, b\}$ from the fundamental representation of $SU(3)_c$. The single-particle space becomes:

$$\mathcal{H}_{\text{quark}}^{(1)} = \mathcal{H}_{\text{scalar}}^{(1)} \otimes \mathcal{H}_{\text{spin}} \otimes \mathcal{H}_{\text{colour}}, \quad (32.5)$$

with $\mathcal{H}_{\text{colour}} \cong \mathbb{C}^3$ and basis vectors:

$$|\mathbf{k}, s, \alpha\rangle, \quad \mathbf{k} \in \mathbb{R}^3, \quad s \in \{1, 2\}, \quad \alpha \in \{r, g, b\}. \quad (32.6)$$

Each further index is a direct product with a new finite-dimensional space. The key distinction is:

	Direct sum $\oplus$	Direct product $\otimes$
What it does	Combines particle-number sectors	Adds a new index to each particle
New basis vecs	From each sector separately	All combinations of both index sets
Dimension	$\dim A + \dim B$	$\dim A \times \dim B$
Example	$\mathcal{H}^{(0)} \oplus \mathcal{H}^{(1)} \oplus \dots$	$\mathcal{H}_{\mathbf{k}} \otimes \mathcal{H}_s \otimes \mathcal{H}_\alpha$
Role in Fock	Organises by particle number	Enriches each particle's label

The direct sum is vertical (across particle numbers); the direct product is horizontal (across index types on a single particle). Every new quantum number we identified in the previous sections—spin, colour, isospin—corresponds to one direct product factor in the single-particle space.

32.2 The full single-particle state

Combining all quantum numbers from the Standard Model fields, the most general single-particle basis state is:

$$|\mathbf{k}, f, s, \alpha, i, \pm\rangle, \quad (32.7)$$

where each label comes from a specific piece of the theory:

Label	Values	Origin	Space
$\mathbf{k}$	$\mathbb{R}^3$ (continuous)	Spacetime translation	$L^2(\mathbb{R}^3)$
f (species/flavour)	$\{e^-, \nu_e, u, d, \dots, \gamma, W, Z, h\}$	Field content of SM	finite
s (spin/helicity)	$\{1, 2\}$ for fermions; $\{\pm 1\}$ for vectors	$SL(2, \mathbb{C})$ rep	$\mathbb{C}^2$
$\pm$ (charge)	particle or antiparticle	Complex field, $b^\dagger$ vs $d^\dagger$	$\mathbb{C}^2$
α (colour)	$\{r, g, b\}$ for quarks; absent for leptons	$SU(3)_c$ fundamental	$\mathbb{C}^3$
i (isospin)	$\{1, 2\}$ for doublets; absent for singlets	$SU(2)_L$ fundamental	$\mathbb{C}^2$

Several labels are absent for certain species. The right-handed electron e_R has no colour index and no isospin index; its single-particle state is just $|\mathbf{k}, e_R, s, \pm\rangle$. The photon has no colour or isospin, and its “spin” becomes the two transverse helicities $\lambda \in \{\pm 1\}$. The left quark doublet Q_L carries all labels: $|\mathbf{k}, Q_L, s, \alpha, i, \pm\rangle$.

The single-particle Hilbert space is the direct product of all contributing spaces:

$$\mathcal{H}_{\text{SM}}^{(1)} = L^2(\mathbb{R}^3) \otimes \mathcal{H}_f \otimes \mathcal{H}_s \otimes \mathcal{H}_\pm \otimes \mathcal{H}_\alpha \otimes \mathcal{H}_i. \quad (32.8)$$

Each tensor factor contributes independently, and operators acting on different factors commute.

32.3 The Standard Model Fock space

One species at a time. Each particle species f has its own Fock space:

$$\mathcal{F}_f = \mathbb{C} \oplus \mathcal{H}_f^{(1)} \oplus \mathcal{H}_f^{(2)} \oplus \dots = \bigoplus_{n=0}^{\infty} \mathcal{H}_f^{(n)}, \quad (32.9)$$

where $\mathcal{H}_f^{(n)}$ is the symmetrised (bosons) or antisymmetrised (fermions) n -fold tensor product of $\mathcal{H}_f^{(1)}$. The direct sum is over particle number for that species alone.

All species together. The full Standard Model Fock space is the direct product of all species-specific Fock spaces:

$$\mathcal{F}_{\text{SM}} = \mathcal{F}_{e^-} \otimes \mathcal{F}_{e^+} \otimes \mathcal{F}_{\nu_e} \otimes \mathcal{F}_\gamma \otimes \mathcal{F}_{u_r} \otimes \mathcal{F}_{u_g} \otimes \mathcal{F}_{u_b} \otimes \dots \otimes \mathcal{F}_h. \quad (32.10)$$

A basis vector of $\mathcal{F}_{\text{SM}}$ specifies the occupation of every mode of every species simultaneously. It is *monosemantic*: every quantum number of every particle is sharp. Schematically:

$$|n_1, \{\mathbf{k}_j^{(1)}, I_j^{(1)}\}; n_2, \{\mathbf{k}_j^{(2)}, I_j^{(2)}\}; \dots\rangle, \quad (32.11)$$

where n_f is the number of particles of species f and $I_j^{(f)} = (s_j, \alpha_j, i_j, \pm_j)$ is the full set of internal quantum numbers of the j -th particle of species f .

Why direct product across species, not direct sum. A physical state can contain electrons *and* photons simultaneously. The state $|1 e^-, 1 \gamma\rangle$ is not a superposition of “one electron, no photons” and “no electrons, one photon”—it genuinely contains both. The direct product allows independent occupation numbers for each species. The direct sum would conflate the two species into a single particle-number count, losing the distinction.

The full structure is therefore: a direct *product* of Fock spaces across species (horizontal, enriching the label), each of which is a direct *sum* over particle numbers for that species (vertical, organising by how many there are).

Equivalent description: enriched index on a single particle type. The decomposition (32.10) into species-specific factors is equivalent to working with a single Fock space whose single-particle states carry a richer label. Instead of separate Fock spaces for u_r, u_g, u_b , one works with a single “quark” Fock space whose single-particle states carry the colour index α . The basis vectors:

$$|n; (\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n)\rangle, \quad I_j = (f_j, s_j, \alpha_j, i_j, \pm_j) \quad (32.12)$$

directly encode all the information of (32.11). The species label f has been absorbed into the index I . Both descriptions refer to the same Hilbert space; the choice between them is a matter of notational convenience.

32.4 Creation and annihilation operators with full indices

For each single-particle state $|\mathbf{k}, I\rangle$ with $I = (f, s, \alpha, i, \pm)$, quantization introduces a pair of operators:

- **Annihilation:** $a(\mathbf{k}, I)$ removes a particle with quantum numbers $(\mathbf{k}, I)$ from the Fock state.
- **Creation:** $a^\dagger(\mathbf{k}, I)$ adds a particle with quantum numbers $(\mathbf{k}, I)$.

For fermions (quarks, leptons) these satisfy the canonical *anticommutation* relations:

$$\{a(\mathbf{k}, I), a^\dagger(\mathbf{k}', I')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{II'}, \quad (32.13)$$

and all other anticommutators vanish. For gauge bosons (photon, W , Z , gluons) and the Higgs, they satisfy commutation relations:

$$[a(\mathbf{k}, I), a^\dagger(\mathbf{k}', I')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{II'}. \quad (32.14)$$

The $\delta_{II'}$ factor is a Kronecker delta over all discrete quantum numbers simultaneously: $\delta_{II'} = \delta_{ff'} \delta_{ss'} \delta_{\alpha\alpha'} \delta_{ii'} \delta_{\pm\pm'}$. Two modes with any differing quantum number are orthogonal and do not mix.

The quantum field operator is the Fourier superposition of all creation and annihilation operators for a given species, generalising the scalar mode expansion:

$$\hat{\Phi}^{s,\alpha,i}(x) = \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[a^{s,\alpha,i}(\mathbf{k}) e^{-ik \cdot x} + b^{s,\alpha,i\dagger}(\mathbf{k}) e^{+ik \cdot x} \right], \quad (32.15)$$

where $a^{s,\alpha,i}(\mathbf{k})$ annihilates a particle and $b^{s,\alpha,i\dagger}(\mathbf{k})$ creates an *antiparticle* with the same quantum numbers. The full index structure (s, α, i) is carried by both the field operator and the creation/annihilation operators: specifying all quantum numbers is what makes the mode expansion unambiguous.

32.5 The Fock hidden state and its evolution

The Standard Model hidden state $h_t \in \mathcal{F}_{\text{SM}}$ is a unit vector expanded in the monosemantic basis (32.12):

$$h_t = \sum_{n=0}^{\infty} \sum_{I_1, \dots, I_n} \int \prod_{j=1}^n \frac{d^3 \mathbf{k}_j}{(2\pi)^3} c(n; (\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n); t) |n; (\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n)\rangle. \quad (32.16)$$

The coefficient $c(\dots; t)$ is the complex amplitude for the configuration of n particles with specified momenta and quantum numbers at time t . The Schrödinger equation:

$$\frac{d h_t}{dt} = -i H_{\text{SM}} h_t \quad (32.17)$$

rotates h_t within $\mathcal{F}_{\text{SM}}$. The Hamiltonian H_{SM} is a sum of interaction terms, each of which is a normal-ordered product of creation and annihilation operators. Each term is an off-diagonal block of H_{SM} in the Fock basis: it transfers amplitude between basis vectors that differ by the particles created or destroyed.

Gauge invariance as a constraint on h_t . Not all vectors in $\mathcal{F}_{\text{SM}}$ are physical. A physical state must be invariant under gauge transformations: in particular, it must be a colour singlet. This means the amplitudes $c(\dots; t)$ are nonzero only on basis vectors where all colour indices are contracted to a singlet. For example:

$$\text{Meson: } |\psi\rangle = \frac{1}{\sqrt{3}} \sum_{\alpha} c(\mathbf{k}_q, \mathbf{k}_{\bar{q}}) |q^{\alpha}(\mathbf{k}_q), \bar{q}_{\alpha}(\mathbf{k}_{\bar{q}})\rangle, \quad (32.18)$$

$$\text{Baryon: } |\psi\rangle = \frac{1}{\sqrt{6}} \sum_{\alpha, \beta, \gamma} \varepsilon_{\alpha\beta\gamma} c(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) |q^{\alpha}(\mathbf{k}_1), q^{\beta}(\mathbf{k}_2), q^{\gamma}(\mathbf{k}_3)\rangle. \quad (32.19)$$

The colour index is contracted (by $\delta^{\alpha\alpha}$ for mesons, by $\varepsilon_{\alpha\beta\gamma}$ for baryons) to give a singlet. The Hamiltonian H_{SM} preserves this constraint: every interaction vertex contracts all colour indices, so the gauge-invariant subspace of $\mathcal{F}_{\text{SM}}$ is invariant under the Schrödinger evolution.

The complete index dictionary for a Fock basis vector.

Label	Values	Origin	Role
n	$0, 1, 2, \dots$	Direct sum	Total particle count
$\mathbf{k}$	$\mathbb{R}^3$ (continuous)	Spacetime translation	3-momentum
f	$e^-, \nu_e, u, d, \dots, \gamma, W, Z, h$	Field content	Species
s	$\{1, 2\}$ or $\{\pm 1\}$	$SL(2, \mathbb{C})$ rep	Spin / helicity
$\pm$	particle / antiparticle	Complex field	Charge sign
α	$\{r, g, b\}$ or absent	$SU(3)_c$ fundamental	Colour
i	$\{1, 2\}$ or absent	$SU(2)_L$ fundamental	Isospin

A single Fock basis vector specifies all of the above for every particle present. Physical states are superpositions of basis vectors in the gauge-invariant subspace. The Standard Model Hamiltonian rotates amplitude among these basis vectors: creating and destroying particles, changing momenta, flipping spins, exchanging colour. Every scattering experiment, every decay, every bound state is a chapter in the story of how H_{SM} moves amplitude through $\mathcal{F}_{\text{SM}}$.

33 Mode Expansion for Gauge Fields

In the scalar field section, the mode expansion was straightforward: one creation/annihilation pair per momentum mode, with the plane wave $e^{i\mathbf{k}\cdot\mathbf{x}}$ as the only basis function. For gauge fields, the mode expansion carries the same logical structure but with two new features. First, each momentum mode has multiple internal slots: spin states, colour states, helicity states. Second, gauge invariance removes some of the naïve degrees of freedom — not all components of a gauge field propagate as physical particles. This section builds the mode expansion for each field type in the Standard Model, connecting the classical multi-index field to the Fock basis vectors.

33.1 Recap: the scalar mode expansion and its logic

For a free real scalar field $\phi(x)$, the mode expansion is:

$$\phi(x) = \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[a(\mathbf{k}) e^{-ik\cdot x} + a^\dagger(\mathbf{k}) e^{+ik\cdot x} \right], \quad (33.1)$$

with $k \cdot x = \omega_{\mathbf{k}}t - \mathbf{k} \cdot \mathbf{x}$ and $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}$. The logic is:

1. The equation of motion (Klein–Gordon: $(\partial^2 + m^2)\phi = 0$) is solved by plane waves $e^{\pm ik\cdot x}$.
2. The general solution is a superposition of all plane waves, labelled by $\mathbf{k}$.
3. Quantization promotes the Fourier coefficients to operators: the positive-frequency coefficient becomes $a(\mathbf{k})$ (annihilation) and the negative-frequency coefficient becomes $a^\dagger(\mathbf{k})$ (creation).
4. The canonical commutation relations $[\phi(x), \pi(y)]_{t=t'} = i\delta^3(\mathbf{x} - \mathbf{y})$ then force $[a(\mathbf{k}), a^\dagger(\mathbf{k}')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}')$.

This same logic applies to every field, but the equation of motion and the plane wave basis are richer: they must account for spin, colour, and gauge constraints.

33.2 Dirac field: spin and particle/antiparticle

The equation of motion. The Dirac equation $(i\gamma^\mu \partial_\mu - m)\Psi = 0$ is solved not by scalar plane waves but by *spinor-valued* plane waves. For each momentum $k^\mu = (\omega_{\mathbf{k}}, \mathbf{k})$ with $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}$, there are four independent solutions:

- Two positive-frequency solutions $u(\mathbf{k}, s) e^{-ik\cdot x}$, $s \in \{1, 2\}$ (particle, two spin states).
- Two negative-frequency solutions $v(\mathbf{k}, s) e^{+ik\cdot x}$, $s \in \{1, 2\}$ (antiparticle, two spin states).

The spinors $u(\mathbf{k}, s) \in \mathbb{C}^4$ and $v(\mathbf{k}, s) \in \mathbb{C}^4$ are fixed 4-component column vectors satisfying:

$$(\not{k} - m) u(\mathbf{k}, s) = 0, \quad (\not{k} + m) v(\mathbf{k}, s) = 0, \quad (33.2)$$

where $\not{k} = \gamma^\mu k_\mu$. They are normalized as:

$$\bar{u}(\mathbf{k}, s) u(\mathbf{k}, s') = 2m \delta_{ss'}, \quad \bar{v}(\mathbf{k}, s) v(\mathbf{k}, s') = -2m \delta_{ss'}, \quad (33.3)$$

and satisfy the completeness (spin sum) relations:

$$\sum_s u^A(\mathbf{k}, s) \bar{u}^B(\mathbf{k}, s) = (\not{k} + m)^{AB}, \quad \sum_s v^A(\mathbf{k}, s) \bar{v}^B(\mathbf{k}, s) = (\not{k} - m)^{AB}, \quad (33.4)$$

where $A, B \in \{1, 2, 3, 4\}$ are Dirac indices. These completeness relations will reappear inside Feynman diagrams when we sum over intermediate spin states.

The mode expansion. Promoting the Fourier coefficients to operators:

$$\Psi^A(x) = \sum_{s=1,2} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[b(\mathbf{k}, s) u^A(\mathbf{k}, s) e^{-ik \cdot x} + d^\dagger(\mathbf{k}, s) v^A(\mathbf{k}, s) e^{+ik \cdot x} \right]. \quad (33.5)$$

Here:

- $b(\mathbf{k}, s)$ annihilates a *particle* (electron) with momentum $\mathbf{k}$ and spin s .
- $d^\dagger(\mathbf{k}, s)$ creates an *antiparticle* (positron) with momentum $\mathbf{k}$ and spin s .
- The spinors u^A and v^A carry the Dirac index A : they encode how the spin label s maps into the four-component structure of Ψ^A .

The anticommutation relations are:

$$\{b(\mathbf{k}, s), b^\dagger(\mathbf{k}', s')\} = \{d(\mathbf{k}, s), d^\dagger(\mathbf{k}', s')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{ss'}. \quad (33.6)$$

All other anticommutators vanish. Fermionic statistics ($b^{\dagger 2} = 0$) enforces Pauli exclusion: at most one fermion per ($\mathbf{k}, s$) mode.

The Dirac index vs. the spin label. It is important to keep these two objects distinct. The Dirac index $A \in \{1, 2, 3, 4\}$ is the row index of the 4×1 column vector Ψ^A ; it is the index acted on by the 4×4 gamma matrices and the Lorentz representation $S(L) = \text{diag}(L, \bar{L})$. The spin label $s \in \{1, 2\}$ is the label on the creation/annihilation operator; it indexes which of the two independent solutions of the Dirac equation at a given $\mathbf{k}$ we are referring to. The spinor $u^A(\mathbf{k}, s)$ is the explicit dictionary between them: given a spin label s and a momentum $\mathbf{k}$, it gives the 4-component amplitude pattern u^A in Dirac space.

Adding colour: the quark field. For a quark field $\Psi^{A,\alpha}(x)$ with colour index $\alpha \in \{r, g, b\}$, the mode expansion carries one additional label:

$$\Psi^{A,\alpha}(x) = \sum_s \sum_\alpha \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[b^\alpha(\mathbf{k}, s) u^A(\mathbf{k}, s) e^{-ik \cdot x} + d^{\alpha\dagger}(\mathbf{k}, s) v^A(\mathbf{k}, s) e^{+ik \cdot x} \right]. \quad (33.7)$$

The colour index α rides on the creation/annihilation operator: $b^\alpha(\mathbf{k}, s)$ annihilates a quark of colour α and spin s at momentum $\mathbf{k}$. The spinors $u^A(\mathbf{k}, s)$ are colour-independent: the Dirac structure does not know about colour. The anticommutation relations become:

$$\{b^\alpha(\mathbf{k}, s), b^{\alpha'\dagger}(\mathbf{k}', s')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{ss'} \delta^{\alpha\alpha'}. \quad (33.8)$$

The $\delta^{\alpha\alpha'}$ factor enforces that operators for different colours commute: a red quark and a blue quark are independent modes. An $SU(3)_c$ gauge transformation mixes these modes:

$$b^\alpha(\mathbf{k}, s) \longmapsto U^\alpha_{\alpha'} b^{\alpha'}(\mathbf{k}, s), \quad (33.9)$$

rotating the colour label exactly as it rotates the field index.

33.3 Gauge boson fields: helicity and the removal of unphysical modes

The naïve count and the problem. A vector field $A^\mu(x)$ has four components $\mu \in \{0, 1, 2, 3\}$. Naïvely, one might expect four independent modes per momentum $\mathbf{k}$. But we showed that gauge invariance removes degrees of freedom: the physical photon has only two propagating modes (the two transverse helicities). The mode expansion must reflect this reduction.

The equation of motion for a free gauge field (Maxwell's equation in covariant form) is $\partial^\nu F_{\mu\nu} = 0$, or equivalently $\partial^2 A_\mu - \partial_\mu(\partial \cdot A) = 0$. In Lorenz gauge $\partial^\mu A_\mu = 0$, this reduces to $\partial^2 A_\mu = 0$: a

massless Klein–Gordon equation for each component. The solutions are plane waves $\epsilon_\mu(\mathbf{k}, \lambda) e^{\pm i\mathbf{k}\cdot\mathbf{x}}$ with $k^2 = 0$ (massless, $\omega_{\mathbf{k}} = |\mathbf{k}|$).

Polarisation vectors. For each momentum $\mathbf{k}$, the *polarisation vectors* $\epsilon_\mu(\mathbf{k}, \lambda)$ are fixed 4-component complex vectors labelled by λ . They are the gauge-boson analogue of the spinors $u^A(\mathbf{k}, s)$: they encode how the helicity label λ maps into the four-component Lorentz vector structure of A^μ . They satisfy:

$$k^\mu \epsilon_\mu(\mathbf{k}, \lambda) = 0 \quad (\text{transversality: Lorenz gauge condition}), \quad (33.10)$$

$$\eta^{\mu\nu} \epsilon_\mu(\mathbf{k}, \lambda) \epsilon_\nu^*(\mathbf{k}, \lambda') = -\delta_{\lambda\lambda'} \quad (\text{normalization}). \quad (33.11)$$

The transversality condition (33.10) eliminates one degree of freedom (the longitudinal component k^μ -direction). A further residual gauge freedom eliminates one more, leaving exactly two independent polarisation vectors per $\mathbf{k}$: the two transverse helicities $\lambda \in \{+1, -1\}$ for a massless vector boson.

The completeness relation for physical polarisations is:

$$\sum_{\lambda=\pm 1} \epsilon_\mu(\mathbf{k}, \lambda) \epsilon_\nu^*(\mathbf{k}, \lambda) = -\eta_{\mu\nu} + \frac{k_\mu \bar{k}_\nu + k_\nu \bar{k}_\mu}{k \cdot \bar{k}}, \quad (33.12)$$

where $\bar{k}^\mu = (\omega_{\mathbf{k}}, -\mathbf{k})$ is an auxiliary lightlike vector. This completeness relation replaces the naïve $-\eta_{\mu\nu}$ and will appear inside loop diagrams when we sum over intermediate photon polarisations.

The photon mode expansion.

$$A^\mu(x) = \sum_{\lambda=\pm 1} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2|\mathbf{k}|}} \left[a(\mathbf{k}, \lambda) \epsilon^\mu(\mathbf{k}, \lambda) e^{-i\mathbf{k}\cdot\mathbf{x}} + a^\dagger(\mathbf{k}, \lambda) \epsilon^{\mu*}(\mathbf{k}, \lambda) e^{+i\mathbf{k}\cdot\mathbf{x}} \right]. \quad (33.13)$$

The operator $a(\mathbf{k}, \lambda)$ annihilates a photon with momentum $\mathbf{k}$ and helicity $\lambda \in \{+1, -1\}$; $a^\dagger(\mathbf{k}, \lambda)$ creates one. The polarisation vector $\epsilon^\mu(\mathbf{k}, \lambda)$ carries the Lorentz index μ and is a fixed classical vector, not an operator. The commutation relations are:

$$[a(\mathbf{k}, \lambda), a^\dagger(\mathbf{k}', \lambda')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{\lambda\lambda'}. \quad (33.14)$$

The Lorentz index vs. the helicity label. As with the Dirac field, we keep two objects distinct. The Lorentz index $\mu \in \{0, 1, 2, 3\}$ is the component index of the 4-vector field A^μ ; it transforms with Λ^μ_ν under Lorentz transformations. The helicity label $\lambda \in \{+1, -1\}$ is the label on the creation/annihilation operator; it specifies which of the two physical polarisation states is being created or destroyed. The polarisation vector $\epsilon^\mu(\mathbf{k}, \lambda)$ is the explicit dictionary between them: given a helicity λ and a momentum $\mathbf{k}$, it gives the 4-component amplitude pattern in Lorentz space.

This is directly parallel to the spinor $u^A(\mathbf{k}, s)$ for the Dirac field: in both cases, a fixed classical object (u^A or ϵ^μ) maps the discrete physical label (s or λ) into the continuous Lorentz representation space ($\mathbb{C}^4$ or $\mathbb{R}^4$).

33.4 Gluon field: colour-adjoint helicity modes

The gluon field $G^{a,\mu}(x)$ carries both a Lorentz index μ (as a vector field) and a colour-adjoint index $a \in \{1, \dots, 8\}$ (from the adjoint representation of $SU(3)_c$). Its mode expansion combines both:

$$G^{a,\mu}(x) = \sum_{\lambda=\pm 1} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2|\mathbf{k}|}} \left[a^a(\mathbf{k}, \lambda) \epsilon^\mu(\mathbf{k}, \lambda) e^{-i\mathbf{k}\cdot\mathbf{x}} + a^{a\dagger}(\mathbf{k}, \lambda) \epsilon^{\mu*}(\mathbf{k}, \lambda) e^{+i\mathbf{k}\cdot\mathbf{x}} \right]. \quad (33.15)$$

The colour label a rides on the creation/annihilation operator, just as the colour label α did for the quark. The polarisation vector $\epsilon^\mu(\mathbf{k}, \lambda)$ is colour-independent: the Lorentz structure does not know about colour. The commutation relations are:

$$[a^a(\mathbf{k}, \lambda), a^{b\dagger}(\mathbf{k}', \lambda')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{\lambda\lambda'} \delta^{ab}. \quad (33.16)$$

The δ^{ab} factor enforces that gluons of different adjoint colour are independent modes. An $SU(3)_c$ gauge transformation acts on the adjoint colour label:

$$a^a(\mathbf{k}, \lambda) \mapsto (Ad_U)^a_b a^b(\mathbf{k}, \lambda), \quad (33.17)$$

rotating the adjoint colour index with the 8×8 adjoint representation matrix.

Why 16 physical gluon modes. At each momentum $\mathbf{k}$, there are $8 \times 2 = 16$ independent gluon modes: 8 colours (adjoint index a) times 2 helicities (helicity label λ). This matches the degree-of-freedom count: $G^{a,\mu}$ has $8 \times 4 = 32$ components, minus 8 from the gauge orbit, minus 8 from transversality, leaving 16 physical modes.

33.5 The W and Z bosons: massive vector fields

The $W^\pm$ and Z bosons acquire mass through the Higgs mechanism, which changes the mode expansion in one important way: a massive vector boson has *three* physical polarisations, not two.

For a massive vector field with mass m_V , the equation of motion is the Proca equation:

$$\partial^\nu F_{\mu\nu} + m_V^2 A_\mu = 0. \quad (33.18)$$

The mass term forces $\partial^\mu A_\mu = 0$ (as a consequence of the equation of motion, not as a gauge choice), eliminating one degree of freedom. But there is no residual gauge freedom to eliminate a second: the mass term breaks the gauge invariance $A_\mu \rightarrow A_\mu + \partial_\mu \alpha$ (since $m_V^2(A_\mu + \partial_\mu \alpha) \neq m_V^2 A_\mu$). The result is $4 - 1 = 3$ physical polarisations: two transverse ($\lambda = \pm 1$) and one longitudinal ($\lambda = 0$).

The longitudinal polarisation vector is:

$$\epsilon^\mu(\mathbf{k}, 0) = \frac{1}{m_V} \left(\frac{|\mathbf{k}|}{m_V}, \frac{\omega_{\mathbf{k}}}{m_V} \hat{\mathbf{k}} \right) \xrightarrow{|\mathbf{k}| \gg m_V} \frac{k^\mu}{m_V}, \quad (33.19)$$

where $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m_V^2}$. At high energies $|\mathbf{k}| \gg m_V$, the longitudinal mode grows as k^μ/m_V : it becomes large, which is the source of the UV divergences in massive vector boson scattering that the Higgs mechanism is designed to cure. The mode expansion for a massive vector boson is:

$$W^{\pm,\mu}(x) = \sum_{\lambda=0,\pm 1} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[a^\pm(\mathbf{k}, \lambda) \epsilon^\mu(\mathbf{k}, \lambda) e^{-ik \cdot x} + a^{\pm\dagger}(\mathbf{k}, \lambda) \epsilon^{\mu*}(\mathbf{k}, \lambda) e^{+ik \cdot x} \right]. \quad (33.20)$$

The sum now runs over three helicities $\lambda \in \{-1, 0, +1\}$. The completeness relation becomes:

$$\sum_{\lambda=0,\pm 1} \epsilon_\mu(\mathbf{k}, \lambda) \epsilon_\nu^*(\mathbf{k}, \lambda) = -\eta_{\mu\nu} + \frac{k_\mu k_\nu}{m_V^2}. \quad (33.21)$$

The extra $k_\mu k_\nu/m_V^2$ term compared to (33.12) is the contribution of the longitudinal mode and grows at high energy, making massive vector boson loops more divergent than photon loops.

33.6 Summary: the mode expansion dictionary

The mode expansion is the precise translation between the classical multi-index field and the Fock basis. The pattern is the same in every case: the equation of motion picks out a set of basis functions (plane waves times a fixed classical object), quantization promotes the Fourier coefficients to creation/annihilation operators, and each discrete label on the operator corresponds to one quantum number on the Fock basis vector.

Field	Spin	Stat.	Classical object	Physical modes per $\mathbf{k}$	Fock label
Scalar ϕ	0	B	1 (trivial)	1	$(\mathbf{k})$
Dirac Ψ^A	$\frac{1}{2}$	F	$u^A(\mathbf{k}, s), v^A(\mathbf{k}, s)$	$2s \times 2$ (p+ap)	$(\mathbf{k}, s, \pm)$
Photon A^μ	1	B	$\epsilon^\mu(\mathbf{k}, \lambda)$	2	$(\mathbf{k}, \lambda)$
$W^\pm, Z$	1	B	$\epsilon^\mu(\mathbf{k}, \lambda)$ (3 pols)	3	$(\mathbf{k}, \lambda), \lambda \in \{-1, 0, +1\}$
Gluon $G^{a,\mu}$	1	B	$\epsilon^\mu(\mathbf{k}, \lambda)$	2×8	$(\mathbf{k}, \lambda, a)$
Quark $\Psi^{A,\alpha}$	$\frac{1}{2}$	F	$u^A(\mathbf{k}, s)$	$2 \times 2 \times 3 \times N_f$	$(\mathbf{k}, s, \pm, \alpha, f)$

In every row, the **classical object** (u^A , ϵ^μ , or 1) is the fixed bridge between the discrete Fock label and the continuous Lorentz representation space. It is determined by the equation of motion and carries no operator content. The **creation/annihilation operators** carry the Fock label and are the operators. The Lagrangian, through the equations of motion, selects which classical objects appear; quantization, through the canonical (anti)commutation relations, promotes the Fourier coefficients to operators.

What the Hamiltonian does to the Fock state. With the mode expansion in hand, the Standard Model Hamiltonian H_{SM} can be expressed entirely in terms of creation and annihilation operators. Each interaction vertex in the Lagrangian — $e\bar{\Psi}\gamma^\mu\Psi A_\mu$ for QED, $g_s\bar{\Psi}^\alpha\gamma^\mu T_{\alpha\beta}^a\Psi^\beta G_\mu^a$ for QCD, and so on — becomes a product of creation and annihilation operators when the mode expansions are substituted in. Each such product maps one Fock basis vector to another:

$$b^\dagger(\mathbf{k}', s') d^\dagger(\mathbf{q}', \lambda') a(\mathbf{p}, \lambda) |\cdots, \gamma(\mathbf{p}, \lambda), \cdots\rangle = |\cdots, e^-(\mathbf{k}', s'), e^+(\mathbf{q}', \lambda'), \cdots\rangle, \quad (33.22)$$

annihilating a photon and creating an electron-positron pair — the elementary amplitude for pair production. The full Schrödinger evolution $\hat{h}_t = -iH_{\text{SM}} h_t$ is the sum of all such elementary transitions, weighted by the coupling constants and phase space factors that come from the mode expansion. This is how the Lagrangian — a sum of index-contracted field products — becomes the Hamiltonian — a sum of creation/annihilation operator products — and how the Hamiltonian becomes the story of particle interactions in the Fock hidden state.

33.7 The spin-statistics theorem: why the choice is forced

The scalar field (Section ??) was quantised with *commutation* relations $[a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'}$. The Dirac field (13.48) was quantised with *anticommutation* relations $\{b_s(\mathbf{p}), b_{s'}^\dagger(\mathbf{p}')\} = \delta_{ss'}\delta_{\mathbf{p}\mathbf{p}'}$. So far, the choice was asserted, not derived. We now prove that the choice is *forced*: the wrong statistics makes the engine unrunnable.

The two choices, the two constraints

When we second-quantise any field, we face a binary choice: *commutation* ($[a, a^\dagger] = 1$, occupation numbers $n = 0, 1, 2, \dots$, Bose–Einstein statistics) or *anticommutation* ($\{b, b^\dagger\} = 1$, occupation

numbers $n = 0, 1$, Fermi–Dirac statistics). A priori, either choice produces a consistent algebra and a valid Fock space. The question is whether the resulting engine runs correctly.

“Runs correctly” means two things:

Constraint 1: stability (energy bounded below). The Hamiltonian H must have a ground state. Without a ground state, the engine can radiate energy indefinitely by creating particles, and the vacuum decays instantly. A runnable engine requires $H \geq E_0$ for some finite E_0 .

Constraint 2: microcausality (no superluminal signaling). A Born-layer measurement at spacetime point x must not affect a Born-layer measurement at a spacelike-separated point y (where $(x - y)^2 < 0$: no light signal can connect them). In the Fock layer, this requires that all local observables built from the field operators commute at spacelike separation:

$$[\hat{O}(x), \hat{O}(y)] = 0 \quad \text{for } (x - y)^2 < 0. \quad (33.23)$$

If this fails, the Born rule allows one observer to send information to another faster than light—not by moving particles, but by choosing which observable to measure.

These two constraints are non-negotiable. A theory that violates either one is not a consistent Fock-layer engine.

What goes wrong: spin- $\frac{1}{2}$ with commutation

Try quantising the Dirac field with commutation relations: $[b_s(\mathbf{p}), b_{s'}^\dagger(\mathbf{p}')] = \delta_{ss'}\delta_{\mathbf{p}\mathbf{p}'}$, and similarly for $d, d^\dagger$.

The second-quantised Hamiltonian, before normal ordering, is:

$$H = \int \frac{d^3p}{(2\pi)^3} E_{\mathbf{p}} \sum_s [b_s^\dagger(\mathbf{p}) b_s(\mathbf{p}) - d_s(\mathbf{p}) d_s^\dagger(\mathbf{p})]. \quad (33.24)$$

The minus sign in front of the d -term comes from the negative-frequency solutions of the Dirac equation: the antiparticle contribution enters with the opposite sign.

With commutation: $d d^\dagger = d^\dagger d + 1$, so:

$$H = \int \frac{d^3p}{(2\pi)^3} E_{\mathbf{p}} \sum_s [\underbrace{b_s^\dagger b_s}_{\geq 0} - \underbrace{d_s^\dagger d_s}_{\geq 0} - 1]. \quad (33.25)$$

The antiparticle number operator $d^\dagger d$ enters with a *negative* sign. Every antiparticle created lowers the energy by $E_{\mathbf{p}}$. With occupation numbers $n = 0, 1, 2, 3, \dots$ (the commutation algebra permits any n), there is no limit: create more and more antiparticles and the energy decreases without bound.

Stability fails. The vacuum is unstable. The engine cannot run.

With anticommutation: $d d^\dagger = 1 - d^\dagger d$, so:

$$H = \int \frac{d^3p}{(2\pi)^3} E_{\mathbf{p}} \sum_s [b_s^\dagger b_s + (1 - d_s^\dagger d_s) - 1] = \int \frac{d^3p}{(2\pi)^3} E_{\mathbf{p}} \sum_s [b_s^\dagger b_s + \bar{d}_s^\dagger \bar{d}_s] - E_0, \quad (33.26)$$

where $\bar{d}^\dagger \equiv d$ reinterprets the antiparticle annihilation as an antiparticle creation (filling the Dirac sea, now implemented algebraically rather than literally) and E_0 is the (infinite but subtractable) vacuum energy. Both terms are positive. Adding any particle or antiparticle raises the energy. The vacuum is stable. The engine runs.

Summary for spin- $\frac{1}{2}$: commutation $\Rightarrow$ energy unbounded below $\Rightarrow$ unstable vacuum. Anticommutation $\Rightarrow$ energy bounded below $\Rightarrow$ stable vacuum.

What goes wrong: spin-0 with anticommutation

Try quantising the scalar field with anticommutation relations: $\{a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger\} = \delta_{\mathbf{k}\mathbf{k}'}$. The Hamiltonian is fine: $H = \sum_{\mathbf{k}} \omega_{\mathbf{k}} a_{\mathbf{k}}^\dagger a_{\mathbf{k}}$, with eigenvalues 0 or $\omega_{\mathbf{k}}$ per mode. Energy bounded below: check.

The problem is microcausality. For any relativistic field, the propagator $\Delta^+(x-y) = \langle 0 | \hat{\phi}(x) \hat{\phi}(y) | 0 \rangle = \int \frac{d^3k}{(2\pi)^3 2\omega_{\mathbf{k}}} e^{-ik \cdot (x-y)}$ does not vanish for spacelike separations. This is unavoidable: Δ^+ is a Lorentz-invariant function that is nonzero even when $(x-y)^2 < 0$. (It falls off exponentially with distance, but it is not zero.)

What saves microcausality is the interplay between particles and antiparticles. For a Hermitian scalar field $\hat{\phi} = \hat{\phi}^{(+)} + \hat{\phi}^{(-)}$ (where $(+)$ contains a and $(-)$ contains $a^\dagger$), the (anti)commutator at spacelike separation is:

$$[\hat{\phi}(x), \hat{\phi}(y)] = \Delta^+(x-y) - \Delta^+(y-x), \quad (33.27)$$

$$\{\hat{\phi}(x), \hat{\phi}(y)\} = \Delta^+(x-y) + \Delta^+(y-x). \quad (33.28)$$

The function $\Delta^+(x-y)$ for spacelike $(x-y)$ satisfies a crucial identity. Since $(x-y)^2 < 0$, there exists a Lorentz frame in which $x^0 = y^0$ (the two events are simultaneous). In this frame, $\Delta^+(x-y)$ depends only on the spatial separation $|\mathbf{x} - \mathbf{y}|$. The same spatial separation with reversed time ordering gives $\Delta^+(y-x)$. Since $x^0 = y^0$, the two are equal:

$$\Delta^+(x-y) = \Delta^+(y-x) \quad \text{for} \quad (x-y)^2 < 0. \quad (33.29)$$

Now read off (33.27) and (33.28):

With commutation: $[\hat{\phi}(x), \hat{\phi}(y)] = \Delta^+ - \Delta^+ = 0$ for spacelike. *Microcausality holds.*

With anticommutation: $\{\hat{\phi}(x), \hat{\phi}(y)\} = \Delta^+ + \Delta^+ = 2\Delta^+ \neq 0$ for spacelike. But $\hat{\phi}$ is Hermitian—it is an observable—so the relevant condition is that its *commutator* vanishes. With anticommuting mode operators, the commutator of $\hat{\phi}$ with itself does not reduce to a c-number; instead, it contains operator-valued terms that do not vanish for spacelike separations. *Microcausality fails.*

(More precisely: with anticommutation, $\hat{\phi}(x)^2$ is not proportional to the identity, and the algebra of local observables built from $\hat{\phi}$ does not close into a set of mutually commuting operators at spacelike separation. The Born-layer measurement at x disturbs the Born-layer measurement at y .)

Summary for spin-0: anticommutation $\Rightarrow$ microcausality fails $\Rightarrow$ superluminal signaling is possible. Commutation $\Rightarrow$ microcausality holds $\Rightarrow$ causality is preserved.

The general case: the role of $(-1)^{2j}$

The two arguments above are instances of a single mechanism. For a field of spin j , the representation theory of the Lorentz group determines the relative phase between the particle and antiparticle contributions to the propagator. Under the exchange $x \leftrightarrow y$ at spacelike separation:

$$\Delta_{\text{antiparticle}}^+(y-x) = (-1)^{2j} \Delta_{\text{particle}}^+(x-y), \quad (33.30)$$

where $(-1)^{2j}$ is the **CPT phase**: +1 for integer spin, -1 for half-integer spin. This factor comes from the transformation properties of the spin- j representation under the combination of charge conjugation, parity, and time reversal.

For the (anti)commutator to vanish at spacelike separation, the particle and antiparticle contributions must cancel:

$$\Delta^+(x-y) + \eta \cdot (-1)^{2j} \Delta^+(x-y) = [1 + \eta \cdot (-1)^{2j}] \Delta^+(x-y) \stackrel{!}{=} 0, \quad (33.31)$$

where $\eta = +1$ for commutation and $\eta = -1$ for anticommutation. For this to vanish:

$$\eta \cdot (-1)^{2j} = -1 \quad \Longleftrightarrow \quad \eta = -(-1)^{2j} = (-1)^{2j+1}. \quad (33.32)$$

Spin j	$(-1)^{2j}$	Required η	Statistics	Consequence
$0, 1, 2, \dots$	$+1$	-1	Commutation (Bose)	$n = 0, 1, 2, \dots$
$\frac{1}{2}, \frac{3}{2}, \dots$	-1	$+1$	Anticommutation (Fermi)	$n = 0, 1$

Integer spin requires commutation. Half-integer spin requires anticommutation. The wrong choice makes the particle and antiparticle contributions *add* instead of *cancel*, and microcausality fails.

Independently, for half-integer spin, commutation also violates stability (the energy is unbounded below, as shown above). Both constraints—stability and microcausality—point to the same conclusion. The pairing is overdetermined: the engine cannot run with the wrong statistics.

The physical picture

The spin-statistics connection has a single physical root: **the exchange of two identical particles is equivalent to rotating one of them by 360°**.

Consider two identical particles at positions $\mathbf{x}_1$ and $\mathbf{x}_2$. Exchange them: move particle 1 to $\mathbf{x}_2$ and particle 2 to $\mathbf{x}_1$. This exchange is topologically equivalent to keeping both particles in place and rotating one of them by 360° around the other (imagine unwinding the exchange path into a single loop).

Under 360° rotation, a spin- j field picks up the phase $(-1)^{2j}$:

$$R(360^\circ) |j, m\rangle = (-1)^{2j} |j, m\rangle. \quad (33.33)$$

For integer spin: $(-1)^{2j} = +1$ (the state is unchanged). For half-integer spin: $(-1)^{2j} = -1$ (the state picks up a sign).

Since exchange = 360° rotation:

$$\hat{\psi}(\mathbf{x}_1)\hat{\psi}(\mathbf{x}_2) = (-1)^{2j} \hat{\psi}(\mathbf{x}_2)\hat{\psi}(\mathbf{x}_1). \quad (33.34)$$

For integer spin: the fields commute $\Rightarrow$ Bose statistics. For half-integer spin: the fields anticommute $\Rightarrow$ Fermi statistics.

This is not a rigorous proof (it assumes that exchange and rotation are topologically equivalent, which requires careful justification in $d \geq 3$ spatial dimensions), but it captures the physical essence: the statistics of identical particles is determined by how a single particle transforms under a full rotation, which is determined by its spin.

The Fock-Born reading

The spin-statistics theorem is a statement about the Fock layer: it constrains the algebra of creation and annihilation operators (commutation vs. anticommutation) and hence the structure of the Fock space (occupation numbers $0, 1, 2, \dots$ vs. $0, 1$). The constraint comes from two Fock-layer requirements (stability and microcausality) and the Lorentz representation (the CPT phase $(-1)^{2j}$).

The Born layer sees the consequences: Bose–Einstein condensation (macroscopic occupation of a single mode, possible only for bosons), the Pauli exclusion principle (at most one fermion per mode), the periodic table (electron shells fill because electrons are fermions), and the stability

of matter (atoms don't collapse because the Pauli pressure of electrons resists gravitational compression).

But the theorem itself lives entirely in the Fock layer. It is a consistency condition on the engine, not a property of the rendering. The engine *cannot run* with the wrong statistics: either the vacuum is unstable (no ground state) or measurements violate causality (superluminal signaling). The spin-statistics pairing is not a law imposed from outside; it is the unique choice that makes the engine self-consistent.

Wrong choice	Fock-layer failure	Born-layer consequence
Spin- $\frac{1}{2}$ + commutation	H unbounded below (vacuum unstable)	Engine radiates infinite energy; no stable matter
Spin-0 + anticommutation	$[\hat{\phi}(x), \hat{\phi}(y)] \neq 0$ for $(x - y)^2 < 0$	Superluminal signaling; Born measurements at spacelike separation interfere

What the reader should take away. The spin-statistics theorem answers the question: “Why does the Dirac field use anticommutation?” Not because we assert it, and not because experiment tells us (though it does). Because the engine has no other choice. Lorentz symmetry fixes the CPT phase $(-1)^{2j}$. Stability and causality together force the statistics. The pairing integer $\leftrightarrow$ Bose, half-integer $\leftrightarrow$ Fermi is a theorem of the Fock layer’s internal consistency.

34 From Lagrangian Core to Hamiltonian Motor: the Schrödinger RNN

The mode expansion of the previous section translated each classical field into a collection of creation and annihilation operators, one pair per momentum mode and per set of internal quantum numbers. We now use this to build the Hamiltonian — the operator that drives the time evolution of the Fock hidden state. The Lagrangian specifies what the interactions are; the Hamiltonian implements them as transitions between Fock basis vectors.

34.1 From Lagrangian to Hamiltonian: the Legendre transform

The four steps. For a field Φ with Lagrangian density $\mathcal{L}(\Phi, \partial_\mu \Phi)$, the Hamiltonian is constructed as follows.

Step 1. Identify the conjugate momentum density:

$$\pi(\mathbf{x}, t) = \frac{\partial \mathcal{L}}{\partial \dot{\Phi}(\mathbf{x}, t)}. \quad (34.1)$$

Step 2. Compute the Hamiltonian density by Legendre transform, trading the time derivative $\dot{\Phi}$ for π :

$$\mathcal{H}(\mathbf{x}, t) = \pi(\mathbf{x}, t) \dot{\Phi}(\mathbf{x}, t) - \mathcal{L}(\mathbf{x}, t). \quad (34.2)$$

The Lagrangian treats $\partial_t \Phi$ and $\partial_i \Phi$ on equal footing. The Hamiltonian density treats them asymmetrically: the time derivative has been traded for π , while the spatial derivatives remain as functions of Φ at a fixed time.

Step 3. Integrate over space to get the Hamiltonian operator:

$$H = \int d^3 \mathbf{x} \mathcal{H}(\mathbf{x}, t). \quad (34.3)$$

Space has been integrated out. What remains is a single operator H acting on Fock space. The spatial structure of the field theory — which modes exist, how they couple — is now encoded in the matrix structure of H , not in an explicit spatial variable.

Step 4. Quantize: promote $\Phi(\mathbf{x})$ and $\pi(\mathbf{x})$ to operators satisfying canonical commutation (for bosons) or anticommutation (for fermions) relations. Substitute the mode expansion for each field. Normal-order the result (place creation operators to the left of annihilation operators) to remove divergent zero-point energies. The result is the Hamiltonian operator H on Fock space — the recurrent weight matrix $-iH$ of the Schrödinger RNN.

Time separated from space. The Legendre transform is precisely the step that separates time from space. The Lagrangian formulation treats all four spacetime coordinates symmetrically and makes Lorentz invariance manifest. The Hamiltonian formulation singles out time: the field theory is now described by a state on a spatial hypersurface at fixed t , evolving forward in time. Lorentz invariance is no longer manifest but remains present: different inertial observers choose different foliations of spacetime, each constructing their own Hamiltonian, and the physical predictions are the same for all.

34.2 The Hamiltonian as a polynomial in creation and annihilation operators

After substituting the mode expansions and normal-ordering, the Hamiltonian splits cleanly into two parts:

$$H = H_{\text{free}} + H_{\text{int}}. \quad (34.4)$$

The free Hamiltonian: diagonal in the Fock basis. The kinetic and mass terms in the Lagrangian give, after mode expansion and normal-ordering:

$$H_{\text{free}} = \sum_{\text{species } f} \sum_{\text{labels } s, \alpha, \dots} \int \frac{d^3 \mathbf{k}}{(2\pi)^3} \omega_{\mathbf{k}}^{(f)} a_f^\dagger(\mathbf{k}, s, \alpha, \dots) a_f(\mathbf{k}, s, \alpha, \dots), \quad (34.5)$$

where $\omega_{\mathbf{k}}^{(f)} = \sqrt{|\mathbf{k}|^2 + m_f^2}$ is the single-particle energy of species f . The number operator $\hat{n}_f(\mathbf{k}) = a_f^\dagger(\mathbf{k}) a_f(\mathbf{k})$ counts particles of species f at momentum $\mathbf{k}$.

Acting on a Fock basis vector, H_{free} returns the same basis vector multiplied by the total energy of the configuration:

$$H_{\text{free}} |(\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n)\rangle = \left(\sum_{j=1}^n \omega_{\mathbf{k}_j}^{(f_j)} \right) |(\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n)\rangle. \quad (34.6)$$

Every Fock basis vector is an eigenstate of H_{free} . The free Hamiltonian is *diagonal* in the Fock basis: it generates no transitions between different particle configurations, only phase rotations within each configuration.

The interaction Hamiltonian: off-diagonal transitions. Each interaction term in the Lagrangian — the QED vertex $e\bar{\Psi}\gamma^\mu\Psi A_\mu$, the quark-gluon vertex $g_s\bar{\Psi}^\alpha\gamma^\mu T_{\alpha\beta}^a\Psi^\beta G_\mu^a$, the Yukawa coupling $y\bar{\Psi}_L\phi\Psi_R$, and so on — contributes a term to H_{int} that is a product of creation and annihilation operators, integrated over momenta with a momentum-conservation delta function:

$$H_{\text{int}} \ni g \int \frac{d^3 \mathbf{k}_1}{(2\pi)^3} \cdots \frac{d^3 \mathbf{k}_r}{(2\pi)^3} (2\pi)^3 \delta^{(3)}\left(\sum_j \epsilon_j \mathbf{k}_j\right) \hat{\mathcal{V}}(a^\dagger, a; \mathbf{k}_1, \dots, \mathbf{k}_r), \quad (34.7)$$

where g is the coupling constant, $\epsilon_j = \pm 1$ for incoming/outgoing momenta, and $\hat{\mathcal{V}}$ is a normal-ordered product of creation and annihilation operators. The delta function is the Fourier transform of the spatial integral $\int d^3 \mathbf{x}$: the Lagrangian density couples fields at the same spacetime point, and this locality produces momentum conservation at each vertex.

Each term in H_{int} is an *off-diagonal block* of H in the Fock basis: it connects basis vectors that differ by the specific particles created and destroyed.

Example. QED: the four elementary vertices.

The QED interaction $\mathcal{L}_{\text{int}} = e\bar{\Psi}\gamma^\mu\Psi A_\mu$ gives a term $H_{\text{int}} \ni e\bar{\Psi}\gamma^\mu\Psi A_\mu$ (with a spatial integral absorbed into a momentum-conservation delta function). Substituting the mode expansions for Ψ , $\bar{\Psi}$, and A_μ :

- Ψ contributes $b(\mathbf{p}, s) + d^\dagger(\mathbf{p}, s)$ (destroy electron or create positron).
- $\bar{\Psi}$ contributes $b^\dagger(\mathbf{p}', s') + d(\mathbf{p}', s')$ (create electron or destroy positron).
- A_μ contributes $a(\mathbf{k}, \lambda) + a^\dagger(\mathbf{k}, \lambda)$ (destroy or create photon).

Multiplying and normal-ordering, the four physically distinct processes are:

Operators	Process	Fock transition
$b^\dagger b a$	Photon absorption	$ e^-, \gamma\rangle \rightarrow e^-\rangle$
$b^\dagger b a^\dagger$	Photon emission	$ e^-\rangle \rightarrow e^-, \gamma\rangle$
$d b a^\dagger$	Pair annihilation	$ e^-, e^+\rangle \rightarrow \gamma\rangle$
$b^\dagger d^\dagger a$	Pair creation	$ \gamma\rangle \rightarrow e^-, e^+\rangle$

Each row is an off-diagonal block of H_{int} mapping one Fock basis vector to another. The coupling constant e sets the amplitude of each transition. The spinors $u^A(\mathbf{k}, s)$, $v^A(\mathbf{k}, s)$ and polarisation vector $\epsilon^\mu(\mathbf{k}, \lambda)$ from the mode expansion encode how the Lorentz indices contract at the vertex.

The general structure.

$$H = \underbrace{H_{\text{free}}}_{\text{diagonal: phase rotation}} + \underbrace{H_{\text{int}}}_{\text{off-diagonal: transitions}}. \quad (34.8)$$

The free part phase-rotates each Fock basis vector at its own energy. The interaction part transfers amplitude between Fock configurations. The Schrödinger RNN runs both simultaneously: rapid phase oscillations within each configuration, and slow redistribution of amplitude between configurations. This redistribution, accumulated over many time steps, is the story of particle physics.

34.3 The Fock vector as a snapshot of the universe

The Fock hidden state $h_t \in \mathcal{F}_{\text{SM}}$ is a unit vector expanded in the monosemantic Fock basis:

$$h_t = \sum_{n=0}^{\infty} \sum_{I_1, \dots, I_n} \int \prod_{j=1}^n \frac{d^3 \mathbf{k}_j}{(2\pi)^3} c(n; (\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n); t) |n; (\mathbf{k}_1, I_1), \dots, (\mathbf{k}_n, I_n)\rangle. \quad (34.9)$$

Time t appears *only in the coefficients* $c(\dots; t)$, not in the basis vectors. The basis vectors are fixed, time-independent objects: they are the possible particle configurations of the universe, labelled by momenta and quantum numbers. The state of the universe at time t is the particular superposition — the particular set of amplitudes $c(\dots; t)$ — that characterises that moment.

The Schrödinger equation:

$$\frac{d}{dt} h_t = -i H_{\text{SM}} h_t, \quad h_t = e^{-i H_{\text{SM}} t} h_0, \quad (34.10)$$

rotates h_t within $\mathcal{F}_{\text{SM}}$. Given the initial state h_0 , the solution is determined for all t . The unitary operator $e^{-i H_{\text{SM}} t}$ rotates the Fock vector; the amplitudes $c(\dots; t)$ evolve accordingly; the basis vectors do not.

The Schrödinger RNN. This is an RNN with the following identification:

RNN concept	QFT concept	Role
Hidden state h_t	Fock vector $h_t \in \mathcal{F}$	Snapshot at time t
Sequence index t	Physical time t	Evolution parameter
Update rule	$e^{-i H dt}$	One Planck step
Recurrent weight	Hamiltonian $-i H$	Fixed matrix
Output	Born rule $ \langle q, h_t \rangle ^2$	Measurement
Hidden state dim	$\dim \mathcal{F} = 2^{10^{183}}$	Astronomically large

The update rule at one Planck step $t_P \approx 5.4 \times 10^{-44}$ s is:

$$h_{t+t_P} = e^{-i H t_P} h_t \approx (I - i H t_P) h_t. \quad (34.11)$$

One matrix-vector multiplication per Planck step. The universe has run approximately 10^{61} such steps since the Big Bang.

Properties of the evolution. The Schrödinger equation (34.10) has four key properties:

- **Linear:** the superposition principle holds exactly. If h_t and h'_t are both solutions, so is any linear combination.
- **Unitary:** $\|h_t\| = \|h_0\| = 1$ for all t . Probabilities are conserved.
- **First-order in t :** the state at $t + dt$ is determined entirely by the state at t . The system is Markovian — it has no memory beyond the current Fock vector.
- **Time-translation invariant:** H does not depend on t (for a time-translation-invariant Lagrangian), so energy is conserved by Noether's theorem.

34.4 Energy, mass, and rotation of the Fock vector

Each basis vector rotates at its own frequency. Under the free Hamiltonian alone, the Schrödinger equation has the exact solution:

$$h_t = e^{-iH_{\text{free}}t} h_0 = \sum_{\mathbf{k}} c(\mathbf{k}) e^{-i\omega_{\mathbf{k}}t} |\mathbf{k}\rangle, \quad (34.12)$$

since each basis vector $|\mathbf{k}\rangle$ is an eigenstate of H_{free} with eigenvalue $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}$. Each basis vector rotates independently in the complex plane at angular frequency $\omega_{\mathbf{k}}$. The amplitude $|c(\mathbf{k})|^2$ for each momentum configuration is constant in time: no transitions occur without interactions.

Mass is the minimum rotation rate. The dispersion relation $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}$ has a minimum at $\mathbf{k} = 0$:

$$\omega_{\mathbf{k}=0} = m. \quad (34.13)$$

A particle at rest rotates at angular frequency m — its rest energy equals its mass in natural units ($c = \hbar = 1$). A massless particle at zero momentum has $\omega = 0$: no rotation, no energy.

Mass is the minimum rate at which a particle basis vector rotates. A heavier particle rotates faster at rest; a lighter one rotates slower. For a multi-particle state, the total energy is the sum of individual energies (in the free theory), so the rotation rate of an n -particle basis vector is $\sum_{j=1}^n \omega_{\mathbf{k}_j}$.

On-shell particles. A Fock basis vector $|\mathbf{k}\rangle$ carries a particle with 3-momentum $\mathbf{k}$ and energy $\omega_{\mathbf{k}}$. The 4-momentum $k^\mu = (\omega_{\mathbf{k}}, \mathbf{k})$ satisfies:

$$k^\mu k_\mu = \omega_{\mathbf{k}}^2 - |\mathbf{k}|^2 = m^2. \quad (34.14)$$

This is the *on-shell condition* $k^2 = m^2$. All Fock basis vectors are on-shell by construction: they are eigenstates of H_{free} with the correct dispersion relation. Real, observable particles — particles that appear in a detector — are always on-shell. They correspond to definite basis vectors in the Fock hidden state.

Virtual particles. When two on-shell particles interact, the perturbative expansion of e^{-iHt} in powers of H_{int} involves intermediate states that do *not* satisfy the on-shell condition. These are *virtual* or *off-shell* particles.

In the Feynman diagram for electron-electron scattering, the exchanged photon carries 4-momentum $q^\mu = p_1^\mu - p'_1{}^\mu$. In general $q^2 = (p_1 - p'_1)^2 \neq 0$: the exchanged photon is off-shell. Its propagator $-i\eta_{\mu\nu}/(q^2 + i\epsilon)$ is finite (no pole), unlike an on-shell photon ($q^2 = 0$, propagator diverges).

Virtual particles do not correspond to Fock basis vectors. They appear in the perturbative expansion of e^{-iHt} , not in the Fock hidden state itself. The exact Fock hidden state at any moment is a superposition of on-shell basis vectors; virtual particles are computational bookkeeping for the perturbative approximation.

	On-shell	Virtual (off-shell)
In Fock basis?	Yes	No
$k^2 = m^2$?	Yes	No
Observable?	Yes (in detector)	No
Where it appears	Fock hidden state	Feynman diagram internal lines
Propagator behaviour	Pole at $q^2 = m^2$	Finite
Physical meaning	Real particle	Perturbative intermediate state

34.5 The vacuum state

The vacuum $|0\rangle$ (or $|\Omega\rangle$ in the interacting theory) is the ground state of the Hamiltonian: the Fock hidden state with the lowest possible energy.

The free vacuum: genuinely empty. For the free Hamiltonian, the vacuum is defined by:

$$a(\mathbf{k}) |0\rangle = 0 \quad \text{for all } \mathbf{k}, \quad (34.15)$$

meaning no particles of any momentum are present. After normal-ordering (which removes the zero-point energy):

$$H_{\text{free}} |0\rangle = 0, \quad e^{-iH_{\text{free}}t} |0\rangle = |0\rangle. \quad (34.16)$$

The free vacuum is stationary: it does not evolve. It is a single Fock basis vector with coefficient 1 and all other coefficients zero. The expected number of particles of any species at any momentum is zero. There are no real particles and no virtual fluctuations: the statement that the vacuum “teems with virtual particles” does not apply to the free vacuum.

The interacting vacuum: shifted and dressed. When H_{int} is turned on, the true ground state $|\Omega\rangle$ of the full Hamiltonian is not the free vacuum $|0\rangle$. The interaction mixes particle-number sectors, so $|\Omega\rangle$ is a superposition:

$$|\Omega\rangle = c_0 |0\rangle + \int \frac{d^3\mathbf{k}}{(2\pi)^3} c_2(\mathbf{k}, -\mathbf{k}) |\mathbf{k}, -\mathbf{k}\rangle + \dots, \quad (34.17)$$

with nonzero amplitude for two-particle, four-particle, and higher-number states (always in momentum-conserving pairs). Nevertheless, $|\Omega\rangle$ is an exact eigenstate of H :

$$H |\Omega\rangle = E_0 |\Omega\rangle, \quad (34.18)$$

so it is stationary and does not evolve dynamically. The interacting vacuum has nonzero particle-number amplitude, but it is a *fixed* quantum state, not a fluctuating one.

The nonzero amplitudes in (34.17) have real physical consequences: the *Lamb shift* (shift in hydrogen energy levels from electron-photon vacuum interactions), the *Casimir effect* (force between conducting plates from photon vacuum fluctuations), and the *anomalous magnetic moment* of the electron (from virtual photon corrections). These are real effects of the multi-particle structure of $|\Omega\rangle$, not of fluctuating virtual particles.

In perturbation theory, $|\Omega\rangle$ is approximated by summing vacuum bubble diagrams — closed Feynman diagrams with no external legs. These are the source of the “teeming virtual particles” description. However, vacuum bubbles cancel in all physical (connected, normalized) amplitudes: they are an artefact of the perturbative approximation and carry no independent physical content.

The Higgs vacuum: a different kind of non-emptiness. In the broken-phase electroweak theory, the physical vacuum satisfies:

$$\langle \Omega | \phi | \Omega \rangle = v \neq 0, \quad (34.19)$$

where $v \approx 246 \text{ GeV}$ is the Higgs vacuum expectation value. This is not a virtual-particle fluctuation. It is a genuine shift of the ground state from $\phi = 0$ to $\phi = v$: a macroscopic classical configuration of the Higgs field filling all of space. Its consequences — fermion masses, W and Z masses — are real and do not cancel in physical amplitudes. The Higgs boson is the small quantum ripple (excitation) on top of this background.

	Free vacuum $ 0\rangle$	Interacting vacuum $ \Omega\rangle$
Definition	$a(\mathbf{k}) 0\rangle = 0$	$H \Omega\rangle = E_0 \Omega\rangle$
Particle content	Exactly zero	Nonzero amplitude for all n
Stationary?	Yes	Yes
Virtual particles?	No	In perturbation theory only
$\langle \phi \rangle$	0	0 (or $v \neq 0$ for Higgs)
Physical effects	None	Lamb shift, Casimir, $g - 2$

The vacuum is the ground state of the Hamiltonian — the lowest-energy Fock vector. Virtual particles are not objects in this state; they are terms in the perturbative expansion of e^{-iHt} , computational artefacts of approximating the exact unitary rotation of the Fock hidden state as a power series in the coupling constants. The exact rotation requires no virtual particles; only the approximation does.

34.6 Fermion and boson fields

We have classified fields by their Lorentz representation (j_L, j_R) . There is a second, independent classification: *statistics*. Two particles of the same species placed in the same state either cancel (fermions) or reinforce (bosons). This is not a separate postulate — it is forced by the combination of Lorentz invariance and the requirement that the theory have a stable ground state. The result is the *spin-statistics theorem*: half-integer spin fields are fermions; integer spin fields are bosons.

The two statistics

Consider two identical particles occupying the same quantum state. Exchanging them returns the system to itself, so the wavefunction can at most pick up a phase:

$$|\mathbf{k}_1, \mathbf{k}_2\rangle = e^{i\theta} |\mathbf{k}_2, \mathbf{k}_1\rangle. \quad (34.20)$$

Exchanging twice must return the original state: $e^{2i\theta} = 1$, so $e^{i\theta} = \pm 1$. The two cases:

- **Bosons:** $e^{i\theta} = +1$. The wavefunction is symmetric under exchange. Multiple particles can occupy the same state. The number of particles in any mode is unconstrained: $n \in \{0, 1, 2, 3, \dots\}$.
- **Fermions:** $e^{i\theta} = -1$. The wavefunction is antisymmetric under exchange. Two particles in the same state give $|\mathbf{k}, \mathbf{k}\rangle = -|\mathbf{k}, \mathbf{k}\rangle$, so $|\mathbf{k}, \mathbf{k}\rangle = 0$: *Pauli exclusion*. At most one particle per mode: $n \in \{0, 1\}$.

In terms of creation and annihilation operators:

$$\text{Bosons: } [a(\mathbf{k}), a^\dagger(\mathbf{k}')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'), \quad (34.21)$$

$$\text{Fermions: } \{b(\mathbf{k}), b^\dagger(\mathbf{k}')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'). \quad (34.22)$$

The bosonic commutator $[a, a^\dagger]$ and the fermionic anticommutator $\{b, b^\dagger\}$ are the algebraic implementations of symmetric and antisymmetric statistics. For fermions, $(b^\dagger)^2 = 0$ from the anticommutator, enforcing Pauli exclusion.

The spin-statistics theorem

The connection between spin and statistics is not an additional assumption — it is a theorem. The proof uses three ingredients: Lorentz invariance, locality (fields at spacelike separation commute or anticommute), and positivity of energy (the Hamiltonian is bounded below). The result:

Lorentz rep	Spin	Statistics	Examples
$(0, 0)$	0	Boson	Higgs h
$(\frac{1}{2}, 0)$ or $(0, \frac{1}{2})$	$\frac{1}{2}$	Fermion	quarks, leptons
$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	$\frac{1}{2}$	Fermion	Dirac spinor Ψ
$(1, \frac{1}{2})$	1	Boson	γ , $W^\pm$, Z^0 , gluons
$(1, 0) \oplus (0, 1)$	1	Boson	(field strength $F_{\mu\nu}$)

Why the wrong statistics leads to disaster. If we tried to quantize a spin- $\frac{1}{2}$ field with bosonic commutators, the Hamiltonian would be unbounded below: for each mode, we could add arbitrarily many fermions to lower the energy without limit, giving an unstable vacuum. If we tried to quantize a spin-1 gauge field with anticommutators, locality would be violated: operators at spacelike separation would fail to commute (or anticommute), allowing faster-than-light signalling. The spin-statistics connection is not a convention — it is forced by consistency.

The ML analogy. In a neural network, different layers can have different activation functions. The choice of activation function for each layer is a design decision that determines the function class the network can represent. Statistics is the analogous design decision for a field: it determines which Fock states are accessible (all $n \geq 0$ for bosons; only $n \in \{0, 1\}$ for fermions) and which algebraic structure (commutator vs anticommutator) governs the creation and annihilation operators. The spin-statistics theorem says this design decision is not free: it is fixed by the Lorentz representation of the field.

Fermion fields: spinors with anticommuting creation operators

A fermion field transforms in a half-integer spin representation of $SL(2, \mathbb{C})$. The two fundamental cases are the Weyl spinors, and their direct sum the Dirac spinor.

Weyl spinors. The left Weyl spinor $\psi_L^s(x)$ ($s \in \{1, 2\}$) and right Weyl spinor $\psi_R^{\dot{s}}(x)$ ($\dot{s} \in \{1, 2\}$) are the minimal fermion fields. Their mode expansions:

$$\psi_L^s(x) = \sum_{s'=1,2} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[b_{s'}(\mathbf{k}) u_L^s(\mathbf{k}, s') e^{-ik \cdot x} + d_{s'}^\dagger(\mathbf{k}) v_L^s(\mathbf{k}, s') e^{+ik \cdot x} \right], \quad (34.23)$$

$$\psi_R^{\dot{s}}(x) = \sum_{s'=1,2} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[b_{s'}(\mathbf{k}) u_R^{\dot{s}}(\mathbf{k}, s') e^{-ik \cdot x} + d_{s'}^\dagger(\mathbf{k}) v_R^{\dot{s}}(\mathbf{k}, s') e^{+ik \cdot x} \right], \quad (34.24)$$

where $u_L^s(\mathbf{k}, s')$ and $u_R^s(\mathbf{k}, s')$ are the two-component spinor wavefunctions for a particle of spin s' and momentum $\mathbf{k}$. The operators $b_{s'}(\mathbf{k})$ and $b_{s'}^\dagger(\mathbf{k})$ annihilate and create fermions; $d_{s'}(\mathbf{k})$ and $d_{s'}^\dagger(\mathbf{k})$ annihilate and create *antifermions*. They satisfy:

$$\{b_s(\mathbf{k}), b_{s'}^\dagger(\mathbf{k}')\} = \{d_s(\mathbf{k}), d_{s'}^\dagger(\mathbf{k}')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{ss'}. \quad (34.25)$$

Particle and antiparticle: two separate Fock spaces. The field $\psi_L^s(x)$ contains both $b^\dagger$ (creating a particle) and d (destroying an antiparticle) because a relativistic field must connect both sectors. This is forced by the requirement that the field transforms locally under Lorentz transformations while also creating positive-energy states. The particle and antiparticle are distinct Fock-space excitations: the electron and positron are both contained in the single Dirac field $\Psi^A(x)$, with the $b^\dagger$ mode creating electrons and the $d^\dagger$ mode creating positrons.

Dirac spinor as a direct sum. The Dirac spinor $\Psi^A(x)$ ($A \in \{1, 2, 3, 4\}$) is the direct sum of a left and right Weyl spinor:

$$\Psi^A(x) = \begin{pmatrix} \psi_L^s(x) \\ \psi_R^s(x) \end{pmatrix}, \quad A = \begin{cases} s \in \{1, 2\} & \text{upper block (left),} \\ s + 2 \in \{3, 4\} & \text{lower block (right).} \end{cases} \quad (34.26)$$

It transforms in the $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation, with $S(L) = \text{diag}(L, \bar{L})$. A Dirac mass term $m\bar{\Psi}\Psi = m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ couples the two Weyl components and is consistent with $SU(3)_c$ gauge invariance (the colour indices contract) but not with $SU(2)_L$ (the isospin indices mismatch).

Majorana fermion. A special case: if the particle is its own antiparticle ($d_s = b_s$), the field is called a *Majorana spinor*. This requires the representation to admit a reality condition compatible with the Lorentz transformation law. For a left Weyl spinor, this is possible if we allow the charge-conjugate of ψ_L to play the role of the right-handed component:

$$\psi_R \equiv i\sigma^2(\psi_L)^*, \quad (34.27)$$

giving a self-conjugate Dirac spinor. Majorana fermions are self-conjugate: they carry no conserved charge (there is no particle/ antiparticle distinction). The neutrino may be a Majorana fermion; whether this is the case is an open experimental question. The Weinberg operator (??) that generates neutrino masses naturally produces Majorana masses.

Boson fields: symmetric creation operators and polarisations

A boson field transforms in an integer spin representation of $SL(2, \mathbb{C})$. The fundamental cases are spin 0 (scalar) and spin 1 (vector).

Scalar field: spin 0, one mode per momentum. The scalar $\phi(x)$ transforms in the $(0, 0)$ representation: no spinor indices, trivial Lorentz transformation $\phi(x) \rightarrow \phi(\Lambda^{-1}x)$. Its mode expansion has one creation operator per momentum mode:

$$\phi(x) = \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2\omega_{\mathbf{k}}}} \left[a(\mathbf{k}) e^{-ik \cdot x} + a^\dagger(\mathbf{k}) e^{+ik \cdot x} \right], \quad (34.28)$$

with $[a(\mathbf{k}), a^\dagger(\mathbf{k}')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}')$. For a real scalar (like the Higgs field h after symmetry breaking), the particle is its own antiparticle: $d^\dagger = a^\dagger$. For a complex scalar, particles and antiparticles are distinct.

Vector field: spin 1 and the polarisation count. The vector field $A^\mu(x)$ transforms in the $(\frac{1}{2}, \frac{1}{2})$ representation, carrying a single Lorentz index $\mu \in \{0, 1, 2, 3\}$. Naively, this gives four degrees of freedom per momentum mode. But the physical degrees of freedom are fewer:

Field	Mass	Physical polarisations	Reduction from
Photon A^μ	0	2 (transverse: $\lambda = \pm 1$)	Gauge inv. (-1) + eom (-1)
$W^\pm, Z^0$	$m_V > 0$	3 (transverse ± 1 + longitudinal 0)	eom only (-1)

For the *massless* photon: the gauge symmetry $A_\mu \rightarrow A_\mu + \partial_\mu \alpha$ removes one degree of freedom (the longitudinal mode), and the equation of motion $k^\mu A_\mu = 0$ (Lorenz condition) removes another, leaving 2 physical transverse helicities $\lambda = \pm 1$. For a *massive* vector (after the Higgs mechanism), gauge invariance is broken, so only the equation of motion $k^\mu A_\mu = 0$ applies (as a consequence of the Proca equation, not a gauge choice), leaving 3 polarisations: $\lambda \in \{-1, 0, +1\}$.

The mode expansion for the photon:

$$A^\mu(x) = \sum_{\lambda=\pm 1} \int \frac{d^3\mathbf{k}}{(2\pi)^3 \sqrt{2|\mathbf{k}|}} \left[a(\mathbf{k}, \lambda) \epsilon^\mu(\mathbf{k}, \lambda) e^{-ik \cdot x} + a^\dagger(\mathbf{k}, \lambda) \epsilon^{\mu*}(\mathbf{k}, \lambda) e^{+ik \cdot x} \right], \quad (34.29)$$

with $[a(\mathbf{k}, \lambda), a^\dagger(\mathbf{k}', \lambda')] = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{\lambda\lambda'}$.

The polarisation vector as a dictionary. The polarisation vector $\epsilon^\mu(\mathbf{k}, \lambda)$ is the analogue of the Dirac spinor $u^A(\mathbf{k}, s)$: a fixed classical object that maps the discrete physical label (λ or s) into the continuous Lorentz representation space ($\mathbb{C}^4$). For the photon:

$$\epsilon^\mu(\mathbf{k}, +1) = \frac{1}{\sqrt{2}}(0, 1, +i, 0), \quad \epsilon^\mu(\mathbf{k}, -1) = \frac{1}{\sqrt{2}}(0, 1, -i, 0), \quad (\text{for } \mathbf{k} = k\hat{z}), \quad (34.30)$$

satisfying $k_\mu \epsilon^\mu = 0$ (transversality) and $\eta_{\mu\nu} \epsilon^\mu \epsilon^{\nu*} = -1$ (normalization).

The Standard Model field content: a unified view

Combining Lorentz representation and statistics, the Standard Model fields are:

Field	Lorentz rep	Spin	Statistics	Physical modes per k	Role
Higgs h	$(0, 0)$	0	Boson	1	Mass generation
Quarks q	$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	$\frac{1}{2}$	Fermion	$2s \times 2$	Matter
Leptons ℓ	$(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$	$\frac{1}{2}$	Fermion	$2s \times 2$	Matter
Photon γ	$(\frac{1}{2}, \frac{1}{2})$	1	Boson	2	EM force
$W^\pm, Z^0$	$(\frac{1}{2}, \frac{1}{2})$	1	Boson	3 (massive)	Weak force
Gluons G^a	$(\frac{1}{2}, \frac{1}{2})$	1	Boson	2×8	Strong force

Two features of this table deserve emphasis.

Matter fields are fermions; force carriers are bosons. This is not a coincidence. Force carriers are gauge bosons: they arise from demanding local gauge invariance, and the gauge field A_μ transforms in the $(\frac{1}{2}, \frac{1}{2})$ representation (integer spin, boson). Matter fields are fermions because the Standard Model is a *chiral* gauge theory: the left- and right-handed components transform differently under $SU(2)_L$, and this chiral structure is possible only for half-integer spin (spinor) representations.

The Higgs is the only fundamental scalar. All other bosons in the Standard Model are spin-1 gauge bosons. The Higgs is the unique spin-0 (scalar) field. This is why it can have a gauge-invariant mass term and a potential $V(\phi)$ with a non-trivial minimum — properties

that are forbidden for gauge bosons. The uniqueness of the Higgs as the fundamental scalar is directly connected to its role as the only species that can generate masses for the others.

Remark. [Supersymmetry: the hypothetical connection] The strict separation between fermions (matter) and bosons (forces) in the Standard Model is one of the motivations for *supersymmetry* (SUSY): a hypothetical symmetry that relates fermion and boson fields by a transformation that changes spin by $\frac{1}{2}$. Under SUSY, each fermion has a bosonic *superpartner* of the same mass and quantum numbers, and vice versa. If SUSY were an exact symmetry, the Higgs mass would be protected from quadratic UV divergences (the boson loop corrections cancel against fermion loop corrections with opposite sign). No superpartner has been observed, so SUSY, if real, must be broken at a scale above the LHC. The spin-statistics theorem still holds: SUSY does not violate it, but rather extends the algebra of symmetry transformations to include generators that change spin.

34.7 QED: the electromagnetic game engine

Quantum electrodynamics is the simplest realistic quantum field theory: it has one gauge symmetry ($U(1)$), two fermion species (electron and positron), and one gauge boson (the photon). Its Hamiltonian is the cleanest illustration of the general structure — free diagonal part plus off-diagonal interaction vertices — and it is the theory whose Hamiltonian formulation was worked out first, by Heisenberg and Pauli in 1929-1930. Every structural feature of the more complicated theories of the Standard Model is present in QED in its simplest form.

The QED Fock space

The fundamental degrees of freedom of QED are:

- *Electrons*: spin- $\frac{1}{2}$ fermions with charge $-e$, mass m_e , and two spin states $s = \pm\frac{1}{2}$. Created by $b^\dagger(\mathbf{k}, s)$, annihilated by $b(\mathbf{k}, s)$, satisfying anticommutation relations $\{b(\mathbf{k}, s), b^\dagger(\mathbf{k}', s')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{ss'}$.
- *Positrons*: the antiparticles of the electron, with charge $+e$, same mass m_e , two spin states. Created by $d^\dagger(\mathbf{k}, s)$, annihilated by $d(\mathbf{k}, s)$.
- *Photons*: spin-1 gauge bosons with charge 0 and mass 0, with two physical transverse polarisations $\lambda = \pm 1$ (in the Coulomb gauge). Created by $a^\dagger(\mathbf{k}, \lambda)$, annihilated by $a(\mathbf{k}, \lambda)$, satisfying bosonic commutation relations.

The QED Fock space is:

$$\mathcal{F}_{\text{QED}} = \mathcal{F}_e^- \otimes \mathcal{F}_e^+ \otimes \mathcal{F}_\gamma, \quad (34.31)$$

where each factor is itself a Fock space: $\mathcal{F}_{e^\pm} = \bigoplus_n \bigwedge^n \mathcal{H}_{e^\pm}^{(1)}$ (antisymmetric, fermionic) and $\mathcal{F}_\gamma = \bigoplus_n \text{Sym}^n \mathcal{H}_\gamma^{(1)}$ (symmetric, bosonic). A general basis vector:

$$|n_e, n_{\bar{e}}, n_\gamma; \{\mathbf{k}_i, s_i\}, \{\mathbf{p}_j, s_j\}, \{\mathbf{q}_l, \lambda_l\}\rangle \quad (34.32)$$

specifies the number and quantum numbers of each species. The Gauss law constraint $\nabla \cdot \hat{\mathbf{E}} = e(\hat{n}_e - \hat{n}_{\bar{e}})$ restricts the physical Hilbert space to states where the longitudinal electric field is sourced by the net charge.

The Coulomb gauge. QED in the Coulomb gauge $\nabla \cdot \mathbf{A} = 0$ eliminates the unphysical longitudinal photon polarisation and the temporal component A_0 . The temporal component is not an independent degree of freedom: it is determined by the Gauss law:

$$-\nabla^2 A_0(\mathbf{x}) = e(\hat{n}_e(\mathbf{x}) - \hat{n}_{\bar{e}}(\mathbf{x})), \quad (34.33)$$

whose solution:

$$A_0(\mathbf{x}) = e \int d^3\mathbf{y} \frac{\hat{n}_e(\mathbf{y}) - \hat{n}_{\bar{e}}(\mathbf{y})}{4\pi|\mathbf{x} - \mathbf{y}|} \quad (34.34)$$

is the instantaneous Coulomb potential generated by the charge density. Substituting (34.34) back into the Hamiltonian produces the Coulomb interaction term explicitly.

The QED Hamiltonian: explicit form

The QED Lagrangian:

$$\mathcal{L}_{\text{QED}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\gamma^\mu D_\mu - m_e)\psi, \quad D_\mu = \partial_\mu + ieA_\mu, \quad (34.35)$$

gives, after the Legendre transform in the Coulomb gauge and quantisation, the QED Hamiltonian:

$$H_{\text{QED}} = H_{\text{free}}^e + H_{\text{free}}^\gamma + H_{\text{Coulomb}} + H_{\text{transverse}}, \quad (34.36)$$

with four terms.

Term 1: free electron-positron Hamiltonian.

$$H_{\text{free}}^e = \int \frac{d^3\mathbf{k}}{(2\pi)^3} E_{\mathbf{k}} \sum_{s=\pm\frac{1}{2}} [b^\dagger(\mathbf{k}, s)b(\mathbf{k}, s) + d^\dagger(\mathbf{k}, s)d(\mathbf{k}, s)], \quad (34.37)$$

where $E_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m_e^2}$. This is diagonal in the Fock basis: each electron and positron mode phase-rotates at its own energy $E_{\mathbf{k}}$. Electrons and positrons have the same mass (Dirac's prediction, confirmed experimentally) and are counted separately by the number operators $b^\dagger b$ and $d^\dagger d$.

Term 2: free photon Hamiltonian.

$$H_{\text{free}}^\gamma = \int \frac{d^3\mathbf{k}}{(2\pi)^3} |\mathbf{k}| \sum_{\lambda=\pm 1} a^\dagger(\mathbf{k}, \lambda)a(\mathbf{k}, \lambda), \quad (34.38)$$

where $|\mathbf{k}|$ is the photon energy (massless: $\omega = |\mathbf{k}|c$) and the sum is over the two physical transverse polarisations. The longitudinal and temporal polarisations are absent in the Coulomb gauge: they have been removed by the gauge constraint, not by the dynamics. The physical Fock space for photons has two, not four, degrees of freedom per momentum mode.

Term 3: the Coulomb interaction.

$$H_{\text{Coulomb}} = \frac{e^2}{2} \int d^3\mathbf{x} d^3\mathbf{y} \frac{\hat{\rho}(\mathbf{x})\hat{\rho}(\mathbf{y})}{4\pi|\mathbf{x} - \mathbf{y}|}, \quad (34.39)$$

where $\hat{\rho}(\mathbf{x}) = \hat{n}_e(\mathbf{x}) - \hat{n}_{\bar{e}}(\mathbf{x})$ is the electric charge density operator. This is an instantaneous, non-local interaction between charged particles: it is the direct consequence of the Gauss law (34.33) and the elimination of A_0 . In momentum space:

$$H_{\text{Coulomb}} = \frac{e^2}{2} \int \frac{d^3\mathbf{q}}{(2\pi)^3} \frac{\hat{\rho}(\mathbf{q})\hat{\rho}(-\mathbf{q})}{|\mathbf{q}|^2}, \quad (34.40)$$

where the $1/|\mathbf{q}|^2$ is the Fourier transform of $1/4\pi r$. This is the Coulomb potential in momentum space: the Born output of (34.39) for two charges is the Coulomb potential $V(r) = -e^2/4\pi r$ (attractive for opposite charges).

The Coulomb term is a four-fermion operator: it contains two factors of $\hat{\rho}$, each quadratic in fermion operators. Explicitly, $\hat{\rho}^2$ contains terms like $b^\dagger b b^\dagger b$ (electron-electron interaction), $d^\dagger d d^\dagger d$ (positron-positron), $b^\dagger b d^\dagger d$ (electron-positron attractive interaction), and so on. It is off-diagonal in the sense that it scatters momenta, but it conserves particle number separately for electrons and positrons.

Term 4: the transverse photon interaction. The coupling of the electron-positron current to the transverse photon field gives the interaction Hamiltonian:

$$H_{\text{transverse}} = e \int d^3\mathbf{x} \, \hat{\bar{\psi}}(\mathbf{x}) \gamma^\mu \hat{\psi}(\mathbf{x}) \hat{A}_\mu^\perp(\mathbf{x}), \quad (34.41)$$

where $\hat{A}_\mu^\perp$ is the transverse photon field (only the two physical polarisations). Substituting the mode expansions for $\hat{\psi}$, $\hat{\bar{\psi}}$, and $\hat{A}_\mu^\perp$, the interaction (34.41) produces four elementary vertex types, as in the example box of the previous section:

Operator structure	Process	Physical content
$b^\dagger(\mathbf{k}', s') b(\mathbf{k}, s) a(\mathbf{q}, \lambda)$	Photon absorption by electron	$e^- + \gamma \rightarrow e^-$
$b^\dagger(\mathbf{k}', s') b(\mathbf{k}, s) a^\dagger(\mathbf{q}, \lambda)$	Photon emission by electron	$e^- \rightarrow e^- + \gamma$
$d(\mathbf{p}', s') b(\mathbf{k}, s) a^\dagger(\mathbf{q}, \lambda)$	Pair annihilation to photon	$e^- + e^+ \rightarrow \gamma$
$b^\dagger(\mathbf{k}, s) d^\dagger(\mathbf{p}, s') a(\mathbf{q}, \lambda)$	Pair creation from photon	$\gamma \rightarrow e^- + e^+$

Each vertex conserves 3-momentum (from the spatial integral in (34.41)) and electric charge (the vertex $e\bar{\psi}\gamma^\mu\psi A_\mu$ is neutral overall: $b^\dagger b$ or $d^\dagger d$ or $b^\dagger d^\dagger$ or bd each conserve charge).

The complete QED vertex factor. Each vertex in (34.41) carries a coupling matrix element from the spinors and polarisation vectors:

$$\mathcal{V}_\mu^{\text{QED}}(k, k', q) = e \bar{u}(k', s') \gamma_\mu u(k, s) \epsilon^\mu(q, \lambda), \quad (34.42)$$

where $u(k, s)$ is the Dirac spinor for the electron at momentum k with spin s , and $\epsilon^\mu(q, \lambda)$ is the photon polarisation vector. The γ^μ matrix encodes how the Lorentz indices contract at the vertex. The vertex factor (34.42) is the explicit matrix element of $H_{\text{transverse}}$ between the initial and final Fock basis vectors, and its modulus squared gives the probability for each transition.

The QED spectrum and its Born outputs

The hydrogen atom as a bound state. The most important Born output of the QED Hamiltonian (34.36) is the hydrogen atom: a bound state of one electron and one proton (treated as a fixed Coulomb source in the non-relativistic limit) with energy spectrum:

$$E_n = m_e + m_p - \frac{\alpha^2 m_e}{2n^2} + O(\alpha^4), \quad (34.43)$$

where $\alpha = e^2/4\pi \approx 1/137$ is the fine structure constant and $n = 1, 2, 3, \dots$ is the principal quantum number. The leading term is the rest energy of the electron and proton; the second term is the binding energy from the Coulomb interaction (34.39); the $O(\alpha^4)$ corrections include the fine structure (spin-orbit coupling from $H_{\text{transverse}}$) and the Lamb shift (from virtual photon corrections to the electron self-energy in $|\Omega\rangle$).

The hydrogen atom is a bound state in the two-particle sector $\mathcal{H}_{e^- p^+}^{(2)}$ of the QED Fock space: its energy $E_n < m_e + m_p$ lies below the two-particle threshold. It is an eigenstate of the full

QED Hamiltonian H_{QED} (or rather, the QED + QCD Hamiltonian that also binds the proton internally) and therefore stationary: the Fock hidden state of the hydrogen atom does not evolve under H_{QED} .

The anomalous magnetic moment. The magnetic moment of the electron, predicted by the Dirac equation to be $g_e = 2$ exactly, receives quantum corrections from virtual photons in the interacting vacuum $|\Omega\rangle$:

$$g_e = 2 \left(1 + \frac{\alpha}{2\pi} + O(\alpha^2) \right), \quad (34.44)$$

where the leading correction $\alpha/2\pi$ is Schwinger's result from the one-loop vertex correction. The experimental value $g_e/2 = 1.00115965218059\dots$ agrees with the theoretical prediction to twelve significant figures: the most precisely verified prediction in the history of physics.

In the Fock-Born framework: the anomalous magnetic moment is the Born output of the magnetic moment observable $\hat{\mu} = -g_e e \hat{\mathbf{S}}/2m_e$ applied to the electron state in the interacting vacuum. The quantum correction $\alpha/2\pi$ arises from the multi-particle components of $|\Omega\rangle$: the physical electron is surrounded by a cloud of virtual photons, and this cloud modifies the effective magnetic moment.

Running coupling: the QED β -function. The fine structure constant α runs with the renormalisation scale μ :

$$\alpha(\mu) = \frac{\alpha(m_e)}{1 - \frac{\alpha(m_e)}{3\pi} \log \frac{\mu^2}{m_e^2}}, \quad (34.45)$$

increasing from $\alpha(m_e) \approx 1/137$ at $\mu = m_e$ to $\alpha(M_Z) \approx 1/129$ at $\mu = M_Z = 91.2 \text{ GeV}$. The QED β -function is positive: $\beta(\alpha) = \alpha^2/3\pi + O(\alpha^3) > 0$, so QED is infrared free (the coupling decreases at long distances).

In the Fock-Born language: the running coupling is the statement that the Born output stability condition requires the coupling α to run with the resolution Λ of the game engine. The β -function encodes the rate of this running.

34.8 Electroweak theory: the symmetry-breaking game engine

The electroweak theory, due to Glashow (1961), Salam (1968), and Weinberg (1967), unifies the electromagnetic and weak interactions into a single $SU(2)_L \times U(1)_Y$ gauge theory. Its Hamiltonian is structurally richer than QED in two ways: the gauge group is non-Abelian ($SU(2)_L$ is non-Abelian, contributing self-interaction vertices among the gauge bosons) and the gauge symmetry is spontaneously broken by the Higgs mechanism, which generates the masses of the $W^\pm$ and Z^0 bosons and all fermions.

The Higgs mechanism is, in the Hamiltonian language, not a choice of Lagrangian but a choice of vacuum: the ground state of the electroweak Hamiltonian is not the state with $\langle\phi\rangle = 0$ but the state with $\langle\phi\rangle = v/\sqrt{2}$. The Fock hidden state of the electroweak vacuum is the state concentrated around this non-zero field configuration, and the $W^\pm$, Z^0 masses are Born outputs of the Hamiltonian evaluated in this vacuum.

The electroweak Fock space

Before symmetry breaking, the fundamental Fock degrees of freedom are the gauge-covariant fields:

- $SU(2)_L$ gauge bosons W_μ^a ($a = 1, 2, 3$): three massless bosons, with creation operators $w^{a\dagger}(\mathbf{k}, \lambda)$ for each polarisation $\lambda = \pm 1, 0$.

- $U(1)_Y$ gauge boson B_μ : one massless boson, with creation operators $b^\dagger(\mathbf{k}, \lambda)$.
- Higgs doublet $\phi = (\phi^+, \phi^0)^T$: a complex scalar $SU(2)_L$ doublet with hypercharge $Y = +1/2$, with four real scalar degrees of freedom (two charged, two neutral).
- Left-handed fermion doublets: $Q_L = (u_L, d_L)^T$, $L_L = (\nu_L, e_L)^T$, and their second and third generation copies.
- Right-handed fermion singlets: u_R , d_R , e_R , and their copies.

The electroweak Fock space before symmetry breaking:

$$\mathcal{F}_{\text{EW}}^{\text{unbroken}} = \mathcal{F}_{W^a} \otimes \mathcal{F}_B \otimes \mathcal{F}_\phi \otimes \mathcal{F}_{\text{fermions}}, \quad (34.46)$$

with 12 gauge boson degrees of freedom (three $SU(2)$ bosons times four polarisations), 4 Higgs degrees of freedom, and the fermionic Fock space for all three generations.

The electroweak Hamiltonian before symmetry breaking

The electroweak Lagrangian is:

$$\begin{aligned} \mathcal{L}_{\text{EW}} = & -\frac{1}{4}W_{\mu\nu}^a W^{a\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \\ & + |D_\mu\phi|^2 - V(\phi) \\ & + \bar{Q}_L i \not{D} Q_L + \bar{L}_L i \not{D} L_L + \bar{u}_R i \not{D} u_R + \bar{d}_R i \not{D} d_R + \bar{e}_R i \not{D} e_R \\ & + \mathcal{L}_{\text{Yukawa}}, \end{aligned} \quad (34.47)$$

where the $SU(2)_L$ field strength:

$$W_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g\epsilon^{abc}W_\mu^b W_\nu^c, \quad (34.48)$$

contains a non-Abelian term $g\epsilon^{abc}W_\mu^b W_\nu^c$ from the $SU(2)$ structure constants ϵ^{abc} . The covariant derivative acting on the Higgs doublet:

$$D_\mu\phi = \left(\partial_\mu - ig\frac{\tau^a}{2}W_\mu^a - ig'\frac{Y}{2}B_\mu \right) \phi, \quad (34.49)$$

where g is the $SU(2)_L$ coupling, g' is the $U(1)_Y$ coupling, τ^a are the Pauli matrices, and $Y = +1/2$ is the Higgs hypercharge.

The Higgs potential:

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4, \quad \mu^2 > 0, \lambda > 0, \quad (34.50)$$

has a Mexican hat shape: the origin $\phi = 0$ is a local maximum (unstable), and the minimum occurs on the circle $|\phi|^2 = \mu^2/2\lambda = v^2/2$ where $v = \mu/\sqrt{\lambda} \approx 246$ GeV.

The Hamiltonian before symmetry breaking. After the Legendre transform, the electroweak Hamiltonian before symmetry breaking contains:

$$\begin{aligned} H_{\text{EW}}^{\text{unbroken}} = & H_{\text{free}}^W + H_{\text{free}}^B + H_{\text{free}}^\phi + H_{\text{free}}^{\text{ferm}} \\ & + H_{\text{gauge-self}}^{SU(2)} + H_{\text{gauge-Higgs}} + H_{\text{gauge-ferm}} + H_{\text{Yukawa}}, \end{aligned} \quad (34.51)$$

where the free terms are diagonal in the Fock basis and the interaction terms are off-diagonal.

The non-Abelian self-interaction of the $SU(2)$ gauge bosons arises from the $W_{\mu\nu}^a W^{a\mu\nu}$ kinetic term. Expanding (34.48):

$$-\frac{1}{4}W_{\mu\nu}^a W^{a\mu\nu} \supset -\frac{g}{2}\epsilon^{abc}(\partial_\mu W_\nu^a - \partial_\nu W_\mu^a)W^{b\mu}W^{c\nu} - \frac{g^2}{4}\epsilon^{abc}\epsilon^{ade}W_\mu^b W_\nu^c W^{d\mu}W^{e\nu}, \quad (34.52)$$

giving:

The three-boson vertex (from the $g\epsilon^{abc}$ term): one vertex with three W bosons, contributing:

$$H_{3W} \ni g\epsilon^{abc} \int d^3\mathbf{x} W_\mu^a (\partial^\mu W^{b\nu} - \partial^\nu W^{b\mu}) W_\nu^c. \quad (34.53)$$

In terms of creation and annihilation operators, this contains terms like $w^{a\dagger} w^{b\dagger} w^c$ (destroy one W , create two), $w^{a\dagger} w^b w^c$ (destroy two W 's, create one), and their Hermitian conjugates. This vertex has no QED analogue: photons do not self-interact because $U(1)$ is Abelian.

The four-boson vertex (from the $g^2\epsilon^{abc}\epsilon^{ade}$ term):

$$H_{4W} \ni \frac{g^2}{4} \epsilon^{abc} \epsilon^{ade} \int d^3\mathbf{x} W_\mu^b W_\nu^c W^{d\mu} W^{e\nu}. \quad (34.54)$$

This contains terms like $w^\dagger w^\dagger w w$ (two-to-two W scattering at order g^2).

The Higgs condensate: spontaneous symmetry breaking at the Hamiltonian level

The Higgs potential (34.50) has its minimum not at $\phi = 0$ but at $|\phi|^2 = v^2/2$. In the Hamiltonian language, this means that the ground state of $H_{\text{EW}}^{\text{unbroken}}$ is not the Fock vacuum $|0\rangle$ (where $\langle\phi\rangle = 0$) but a different state $|\Omega\rangle_{\text{EW}}$ where $\langle\phi\rangle \neq 0$.

Choosing the vacuum. By the $SU(2)_L \times U(1)_Y$ gauge freedom, we can always rotate the vacuum expectation value to the form:

$$\langle\Omega_{\text{EW}}|\hat{\phi}|\Omega_{\text{EW}}\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad v \approx 246 \text{ GeV}. \quad (34.55)$$

This is the unitary gauge choice: the three Goldstone bosons (the three would-be Goldstone fields corresponding to the broken $SU(2)_L \times U(1)_Y \rightarrow U(1)_{\text{EM}}$ generators) are “eaten” by the W_μ^1 , W_μ^2 , and Z_μ^0 gauge bosons, which acquire longitudinal polarisations and become massive.

In the Fock-Born framework: the vacuum (34.55) is a specific amplitude in the Higgs Fock space. The condensate $\langle\phi\rangle = v/\sqrt{2}$ is the Born output of the Higgs field operator in the ground state — the expectation value of $\hat{\phi}^0$ in the interacting vacuum $|\Omega_{\text{EW}}\rangle$. It is a Level 4 Born output (a macroscopic classical field configuration) of the Fock hidden state. The Goldstone bosons that are eaten are not physical particles; they are degrees of freedom of the Higgs Fock space that become the longitudinal polarisations of the massive gauge bosons in the unitary gauge.

Expanding around the vacuum. Writing $\phi = (v + h)/\sqrt{2}$ (in the unitary gauge where the Goldstone bosons are eliminated), the Higgs field h is the physical Higgs boson: the small oscillation around the vacuum. Substituting into the potential:

$$V(v + h) = -\frac{\mu^2}{2}(v + h)^2 + \frac{\lambda}{4}(v + h)^4 = \text{const} + \frac{1}{2}(2\mu^2)h^2 + \sqrt{\frac{\lambda}{2}}\mu h^3 + \frac{\lambda}{4}h^4, \quad (34.56)$$

giving the Higgs boson mass:

$$m_H^2 = 2\mu^2 = 2\lambda v^2. \quad (34.57)$$

The experimental value is $m_H = 125.25 \text{ GeV}$, giving $\lambda = m_H^2/2v^2 \approx 0.129$.

The electroweak Hamiltonian after symmetry breaking: mass generation

Substituting $\phi = (0, (v + h)/\sqrt{2})^T$ into the Higgs kinetic term $|D_\mu\phi|^2$ generates mass terms for the gauge bosons. The covariant derivative (34.49) acting on the vacuum:

$$D_\mu\langle\phi\rangle = -\frac{i}{\sqrt{2}} \begin{pmatrix} \frac{g}{2}(W_\mu^1 - iW_\mu^2)v \\ (-\frac{g}{2}W_\mu^3 + \frac{g'}{2}B_\mu)v \end{pmatrix}, \quad (34.58)$$

and $|D_\mu\langle\phi\rangle|^2$ gives the gauge boson mass terms.

Mass eigenstates: $W^\pm$, Z^0 , and γ . The physical gauge bosons are the mass eigenstates:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), \quad (34.59)$$

$$Z_\mu^0 = \cos\theta_W W_\mu^3 - \sin\theta_W B_\mu, \quad (34.60)$$

$$A_\mu = \sin\theta_W W_\mu^3 + \cos\theta_W B_\mu, \quad (34.61)$$

where the *Weinberg angle* θ_W satisfies:

$$\tan\theta_W = \frac{g'}{g}, \quad \sin^2\theta_W \approx 0.231. \quad (34.62)$$

The photon A_μ is the massless combination (the unbroken $U(1)_{\text{EM}}$ gauge boson); the Z^0 is the massive neutral combination; the $W^\pm$ are the massive charged bosons.

The mass terms arising from $|D_\mu\phi|^2$:

$$m_W^2 W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} Z_\mu^0 Z^{0\mu}, \quad (34.63)$$

give the gauge boson masses:

$$m_W = \frac{gv}{2}, \quad m_Z = \frac{\sqrt{g^2 + g'^2}v}{2} = \frac{m_W}{\cos\theta_W}. \quad (34.64)$$

With $v = 246$ GeV, $g \approx 0.653$, $g' \approx 0.357$: $m_W \approx 80.4$ GeV and $m_Z \approx 91.2$ GeV, in agreement with the experimental values.

In the Fock-Born framework: the masses (34.64) are Born outputs of the electroweak Hamiltonian evaluated in the symmetry-broken vacuum (34.55). They are not inputs to the Lagrangian; they are derived from the Higgs condensate v , the $SU(2)$ coupling g , and the $U(1)_Y$ coupling g' . Before the condensate forms ($v = 0$, at temperatures above the electroweak phase transition $T_c \approx 160$ GeV), the game engine has massless W^a and B bosons; after the condensate forms ($v \neq 0$), the game engine has massive $W^\pm$ and Z^0 and a massless photon. The phase transition is a change in the ground state of the game engine, not a change in the game engine itself.

The Higgs-gauge boson interactions. Expanding $|D_\mu\phi|^2$ with $\phi = (0, (v + h)/\sqrt{2})^T$ also produces Higgs-gauge boson interaction vertices:

$$H_{\text{H-gauge}} \ni \frac{g^2 v}{2} h W_\mu^+ W^{-\mu} + \frac{g^2}{4} h^2 W_\mu^+ W^{-\mu} + \frac{(g^2 + g'^2)v}{4} h Z_\mu^0 Z^{0\mu} + \dots, \quad (34.65)$$

giving the HWW and HZZ vertices with coupling strengths $g^2 v/2$ and $(g^2 + g'^2)v/4$ proportional to the gauge boson masses. These vertices are directly measured at the LHC through Higgs production and decay processes.

The broken-phase electroweak Fock space. After symmetry breaking, the physical Fock space reorganises:

Field	Mass	Charge	Fock operators
Photon A_μ	0	0	$a^\dagger(\mathbf{k}, \lambda), \lambda = \pm 1$
W_μ^+	80.4 GeV	$+e$	$w^{+\dagger}(\mathbf{k}, \lambda), \lambda = 0, \pm 1$
W_μ^-	80.4 GeV	$-e$	$w^{-\dagger}(\mathbf{k}, \lambda), \lambda = 0, \pm 1$
Z_μ^0	91.2 GeV	0	$z^\dagger(\mathbf{k}, \lambda), \lambda = 0, \pm 1$
Higgs h	125.25 GeV	0	$h^\dagger(\mathbf{k})$

The three Goldstone bosons that were part of $\mathcal{F}_\phi^{\text{unbroken}}$ have become the longitudinal polarisations $\lambda = 0$ of the $W^\pm$ and Z^0 : three new Fock degrees of freedom that were absent for massless gauge bosons. This is the Higgs mechanism at the Fock space level: the symmetry breaking reorganises the Fock degrees of freedom, not just the mass spectrum.

The electroweak vacuum: structure and stability

The electroweak vacuum $|\Omega_{\text{EW}}\rangle$ is the ground state of $H_{\text{EW}}^{\text{broken}}$. Its structure in the Fock-Born framework:

Higgs condensate: the Born output $\langle \hat{\phi}^0 \rangle = v/\sqrt{2}$ is a Level 4 Born output — a macroscopic classical field filling all of space. It is not a particle; it is the classical value of the Higgs field in the ground state.

Vacuum stability: the top Yukawa coupling drives the quartic Higgs coupling λ_H negative at scales $\sim 10^{10}$ – 10^{12} GeV (Section ??), making the electroweak vacuum metastable. The tunnelling rate to the true vacuum is extremely small (the universe lifetime is much shorter than the tunnelling time), but the metastability implies that the electroweak vacuum is not the true ground state.

The hierarchy problem: the Higgs mass squared is a relevant operator at the Gaussian fixed point with $y_{m_H^2} = 2$, driven to $m_H^2 \sim \Lambda^2$ by the RG flow. The observed $m_H = 125$ GeV requires fine-tuning of order $(m_H/M_{\text{Pl}})^2 \sim 10^{-34}$ if no new physics intervenes below the Planck scale. In the Fock-Born language: the Born output stability condition for m_H^2 requires an extraordinary cancellation between the bare Higgs mass and the quantum corrections from the top quark loop, the $W^\pm$ loop, and the Higgs self-energy.

34.9 QCD: the strong game engine

Quantum chromodynamics is the $SU(3)$ gauge theory of the strong interaction. Its Hamiltonian has the same structure as the electroweak Hamiltonian — free diagonal terms plus off-diagonal interaction vertices — but with two crucial differences: the gauge group is $SU(3)$ (with $N_c^2 - 1 = 8$ generators) rather than $SU(2)$, and the gauge symmetry is *unbroken* (no Higgs condensate in the QCD sector). The unbroken non-Abelian gauge symmetry leads to the most striking qualitative difference between QCD and QED: colour confinement, the fact that no isolated quark or gluon appears as a physical particle.

The QCD Fock space

The fundamental Fock degrees of freedom of QCD are:

- *Quarks:* spin- $\frac{1}{2}$ fermions carrying colour charge $\alpha = 1, 2, 3$ (red, green, blue) and flavour $f = u, d, s, c, b, t$. Created by $q_\alpha^{f\dagger}(\mathbf{k}, s)$, with anticommutation relations: $\{q_\alpha^f(\mathbf{k}, s), q_\beta^{f'\dagger}(\mathbf{k}', s')\} = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \delta_{\alpha\beta} \delta_{ff'} \delta_{ss'}$.

- *Antiquarks*: $\bar{q}_\alpha^{f\dagger}(\mathbf{k}, s)$, with colour in the conjugate representation $\bar{\mathbf{3}}$.
- *Gluons*: spin-1 gauge bosons carrying colour in the adjoint representation of $SU(3)$ (colour index $a = 1, \dots, 8$), massless, with two physical polarisations in the Coulomb gauge. Created by $g^{a\dagger}(\mathbf{k}, \lambda)$.

The QCD Fock space:

$$\mathcal{F}_{\text{QCD}} = \mathcal{F}_{\text{quark}} \otimes \mathcal{F}_{\text{antiquark}} \otimes \mathcal{F}_{\text{gluon}}. \quad (34.66)$$

The Gauss law constraint:

$$\hat{G}^a(\mathbf{x}) |\text{phys}\rangle = 0, \quad \hat{G}^a = \partial_i E^{ai} + g_s f^{abc} A_i^b E^{ci} - g_s \sum_f q_f^\dagger T^a q_f \quad (34.67)$$

restricts the physical Hilbert space to colour-singlet states. Here E^{ai} is the chromoelectric field operator (conjugate to A_i^a), f^{abc} are the $SU(3)$ structure constants, and T^a are the generators in the fundamental representation. No free quark or gluon satisfies (34.67): both carry colour charge and are not colour-singlet. The physical Fock space consists only of colour-singlet combinations.

The QCD Hamiltonian: explicit form

The QCD Lagrangian:

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} G_{\mu\nu}^a G^{a\mu\nu} + \sum_{f=u,d,s,c,b,t} \bar{q}_f (i\gamma^\mu D_\mu - m_f) q_f, \quad (34.68)$$

where the gluon field strength:

$$G_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g_s f^{abc} A_\mu^b A_\nu^c, \quad (34.69)$$

and the covariant derivative $D_\mu q_f = (\partial_\mu - ig_s T^a A_\mu^a) q_f$.

After the Legendre transform in the Coulomb gauge (where $\partial_i A^{ai} = 0$), the QCD Hamiltonian is:

$$H_{\text{QCD}} = H_{\text{free}}^q + H_{\text{free}}^g + H_{\text{Coulomb}}^{\text{QCD}} + H_{\text{quark-gluon}} + H_{3g} + H_{4g}, \quad (34.70)$$

with six terms.

Term 1: free quark Hamiltonian.

$$H_{\text{free}}^q = \sum_f \int \frac{d^3\mathbf{k}}{(2\pi)^3} E_{\mathbf{k}}^f \sum_{\alpha=1}^3 \sum_{s=\pm\frac{1}{2}} [q_\alpha^{f\dagger}(\mathbf{k}, s) q_\alpha^f(\mathbf{k}, s) + \bar{q}_\alpha^{f\dagger}(\mathbf{k}, s) \bar{q}_\alpha^f(\mathbf{k}, s)], \quad (34.71)$$

where $E_{\mathbf{k}}^f = \sqrt{|\mathbf{k}|^2 + m_f^2}$ and the sum is over six flavours, three colours, and two spins. The quark creation operators carry a colour index α that transforms in the fundamental representation $\mathbf{3}$ of $SU(3)$.

Term 2: free gluon Hamiltonian.

$$H_{\text{free}}^g = \int \frac{d^3\mathbf{k}}{(2\pi)^3} |\mathbf{k}| \sum_{a=1}^8 \sum_{\lambda=\pm 1} g^{a\dagger}(\mathbf{k}, \lambda) g^a(\mathbf{k}, \lambda), \quad (34.72)$$

with 8 colour labels and 2 physical polarisations per mode. The gluon carries colour in the adjoint representation $\mathbf{8}$ of $SU(3)$: this is the root cause of the non-Abelian self-interaction.

Term 3: the QCD Coulomb interaction.

$$H_{\text{Coulomb}}^{\text{QCD}} = \frac{g_s^2}{2} \int d^3\mathbf{x} d^3\mathbf{y} \hat{\rho}^a(\mathbf{x}) \frac{-1}{\nabla^2 + g_s^2 \hat{f}_{\text{adj}}}{}^{ab}(\mathbf{x}, \mathbf{y}) \hat{\rho}^b(\mathbf{y}), \quad (34.73)$$

where $\hat{\rho}^a = q^\dagger T^a q + f^{abc} A_i^b E^{ci}$ is the colour charge density (quarks plus gluons) and the kernel includes a Gribov-Zwanziger factor from the non-Abelian Gauss law constraint. In the weak-coupling limit, this reduces to the colour-Coulomb interaction between quarks:

$$H_{\text{Coulomb}}^{\text{QCD}} \xrightarrow{g_s \rightarrow 0} -\frac{g_s^2}{2} \int \frac{d^3\mathbf{q}}{(2\pi)^3} \frac{\hat{\rho}^a(\mathbf{q}) \hat{\rho}^a(-\mathbf{q})}{|\mathbf{q}|^2}. \quad (34.74)$$

The colour factor: for a quark-antiquark pair in the colour-singlet state (the meson), $\langle T^a \otimes \bar{T}^a \rangle = -C_F = -4/3$ (the fundamental Casimir). This gives an attractive short-distance Coulomb potential $V(r) = -4\alpha_s/3r$, reproducing the first term of the Cornell potential (??).

Term 4: the quark-gluon vertex.

$$H_{\text{quark-gluon}} = g_s \int d^3\mathbf{x} \hat{q}_f(\mathbf{x}) \gamma^\mu T^a \hat{q}_f(\mathbf{x}) \hat{A}_\mu^a(\mathbf{x}), \quad (34.75)$$

the QCD analogue of the QED vertex (34.41). The colour matrix T^a (a 3×3 matrix in colour space) appears at each vertex, giving a colour factor that depends on the colour state of the quark. The vertex processes are:

Process	Fock transition	Colour structure
Gluon emission by quark	$q_\alpha^f \rightarrow q_\beta^f + g^a$	$T_{\beta\alpha}^a$ (colour change by T^a)
Gluon absorption by quark	$q_\alpha^f + g^a \rightarrow q_\beta^f$	$T_{\beta\alpha}^a$
Quark pair creation	$g^a \rightarrow q_\alpha^f + \bar{q}_\beta^f$	$T_{\alpha\beta}^a$
Quark pair annihilation	$q_\alpha^f + \bar{q}_\beta^f \rightarrow g^a$	$T_{\alpha\beta}^a$

The key difference from QED: the colour matrix T^a changes the colour of the quark at each vertex. A quark that emits a gluon changes its colour (a red quark can emit a gluon and become a blue quark, for instance). The colour is not destroyed but transferred to the gluon.

Term 5: the three-gluon vertex. The non-Abelian term $f^{abc} A_\mu^b A_\nu^c$ in the gluon field strength (34.69) produces the three-gluon vertex:

$$\begin{aligned} H_{3g} &\ni g_s f^{abc} \int d^3\mathbf{x} (\partial_\mu A_\nu^a - \partial_\nu A_\mu^a) A^{b\mu} A^{c\nu} \\ &= g_s f^{abc} \int \frac{d^3\mathbf{k}_1 d^3\mathbf{k}_2 d^3\mathbf{k}_3}{(2\pi)^9} (2\pi)^3 \delta^3(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) \mathcal{V}_{\mu\nu\rho}^{abc}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) A^{a\mu} A^{b\nu} A^{c\rho}. \end{aligned} \quad (34.76)$$

The tensor structure $\mathcal{V}_{\mu\nu\rho}^{abc}$ is:

$$\mathcal{V}_{\mu\nu\rho}^{abc}(k_1, k_2, k_3) = f^{abc} [g_{\mu\nu}(k_1 - k_2)_\rho + g_{\nu\rho}(k_2 - k_3)_\mu + g_{\rho\mu}(k_3 - k_1)_\nu]. \quad (34.77)$$

In the Fock basis, H_{3g} contains terms like $g^{a\dagger} g^{b\dagger} g^c$ (destroy one gluon, create two) and $g^{a\dagger} g^b g^c$ (destroy two gluons, create one). The structure constant f^{abc} encodes how the colour of the three gluons must combine: the three-gluon vertex conserves colour in the sense that the colour indices must contract into a colour-singlet via f^{abc} .

Term 6: the four-gluon vertex. The quadratic non-Abelian term in $G_{\mu\nu}^a G^{a\mu\nu}$:

$$H_{4g} \ni \frac{g_s^2}{4} f^{abc} f^{ade} \int d^3\mathbf{x} A_\mu^b A_\nu^c A^{d\mu} A^{e\nu}, \quad (34.78)$$

contains terms like $g^\dagger g^\dagger gg$ (two-to-two gluon scattering at order g_s^2). The colour structure $f^{abc} f^{ade}$ is the product of two structure constants, giving a more complex colour factor than the three-gluon vertex.

The QCD colour factor hierarchy. The various QCD processes have different colour factors determined by the $SU(3)$ Casimirs:

Process	Colour factor	Origin
Quark-antiquark (singlet) attractive Coulomb	$C_F = \frac{N_c^2 - 1}{2N_c} = \frac{4}{3}$	Fundamental Casimir
Gluon self-energy	$C_A = N_c = 3$	Adjoint Casimir
Quark loop in gluon self-energy	$T_F = \frac{1}{2}$	Trace normalisation
Quark-quark (triplet) repulsive	$\frac{C_F}{2} = \frac{2}{3}$	Product of colour factors
Quark-quark (sextet) attractive	$-\frac{C_F}{4} = -\frac{1}{3}$	Colour arithmetic

The colour factor $C_F = 4/3$ for the quark-antiquark singlet (the meson configuration) is attractive: this is why mesons are bound states. The colour factor for quark-quark scattering in the antisymmetric (antitriplet) channel is also attractive, producing the diquark correlation that contributes to baryon binding.

Confinement: the physical Hilbert space and the string tension

The most striking feature of the QCD Hamiltonian is that it does not support free quark or gluon states in the physical Hilbert space. The Gauss law (34.67) requires all physical states to be colour-singlets. This has two consequences.

The colour-singlet constraint. A single quark state $q_\alpha^{f\dagger}(\mathbf{k}, s) |0\rangle$ is not colour-singlet: it transforms as a colour triplet $\mathbf{3}$ under $SU(3)$. To satisfy the Gauss law, it must be combined with other colour charges to form a singlet:

$$\text{Meson : } \frac{1}{\sqrt{3}} \sum_\alpha q_\alpha^{f\dagger}(\mathbf{k}, s) \bar{q}_\alpha^{f'\dagger}(\mathbf{p}, s') |\Omega\rangle \in \mathcal{H}^{\text{phys}}, \quad (34.79)$$

$$\text{Baryon : } \frac{1}{\sqrt{6}} \epsilon^{\alpha\beta\gamma} q_\alpha^{f\dagger} q_\beta^{f'\dagger} q_\gamma^{f''\dagger} |\Omega\rangle \in \mathcal{H}^{\text{phys}}, \quad (34.80)$$

$$\text{Free quark : } q_\alpha^{f\dagger}(\mathbf{k}, s) |\Omega\rangle \notin \mathcal{H}^{\text{phys}}. \quad (34.81)$$

The $\epsilon^{\alpha\beta\gamma}$ in (34.80) is the totally antisymmetric tensor that contracts three colour indices into a singlet. The single quark state (34.81) violates the Gauss law and is not in the physical Hilbert space.

Confinement from the string tension. Beyond the gauge constraint, the dynamics of the QCD Hamiltonian produce a confining potential: as two colour-charged objects (a quark and antiquark in a meson) are separated, the energy of the chromoelectric field between them grows linearly with the separation:

$$V(r) = \sigma r - \frac{4\alpha_s}{3r} + C + O(r^{-2}), \quad (34.82)$$

where $\sigma \approx 0.18 \text{ GeV}^2 \approx 1 \text{ GeV/fm}$ is the string tension (Section ??). The linear term grows without bound: the energy required to separate the quark and antiquark to infinity is infinite.

In the Fock-Born framework: the string tension is a Born output of the QCD Hamiltonian. It is the rate at which the energy eigenvalue of the quark-antiquark bound state grows with the separation, computed from the expectation value of H_{QCD} in the flux tube state. The confining potential (34.82) is derived by the Feshbach-Löwdin projection that integrates out the short-distance QCD degrees of freedom (high-momentum quarks and gluons) and retains only the long-distance flux tube collective mode.

Asymptotic freedom: the running of α_s . The QCD β -function (Section ??):

$$\beta(g_s) = -\frac{g_s^3}{16\pi^2} \left(\frac{11N_c}{3} - \frac{2N_f}{3} \right) + O(g_s^5) \quad (34.83)$$

is negative for $N_f < 11N_c/2 = 16.5$: with $N_c = 3$ and $N_f = 6$ active flavours, $\beta(g_s) < 0$ and the coupling $\alpha_s = g_s^2/4\pi$ decreases at shorter distances.

$$\alpha_s(\mu) \approx \frac{2\pi}{b_0 \log(\mu/\Lambda_{\text{QCD}})}, \quad b_0 = 11 - \frac{2N_f}{3} \approx 7, \quad (34.84)$$

where $\Lambda_{\text{QCD}} \approx 200 \text{ MeV}$ is the non-perturbative scale where $\alpha_s \sim 1$ and confinement sets in.

For $\mu \gg \Lambda_{\text{QCD}}$: $\alpha_s \ll 1$ and the quark-gluon interaction is weak; the Dyson series converges; quarks and gluons behave approximately as free particles (asymptotic freedom). For $\mu \lesssim \Lambda_{\text{QCD}}$: $\alpha_s \sim 1$ and the perturbative expansion breaks down; the full non-perturbative Hamiltonian must be used; confinement and hadronisation occur.

The transition from the perturbative regime ($\mu > \Lambda_{\text{QCD}}$, free quarks and gluons) to the non-perturbative regime ($\mu < \Lambda_{\text{QCD}}$, confined hadrons) is the Feshbach-Löwdin projection that produces the Cornell potential (34.82) from the QCD Hamiltonian: the high-energy quark and gluon degrees of freedom are integrated out, producing the effective quark-antiquark potential.

The QCD vacuum: a condensate-filled game engine

Unlike the electroweak vacuum, the QCD vacuum does not involve a classically displaced field configuration ($\langle A_\mu^a \rangle = 0$ by Lorentz invariance). Instead, the QCD vacuum is characterised by non-zero vacuum condensates: non-zero expectation values of gauge-invariant operators.

The chiral condensate. In the limit of massless u and d quarks, the QCD Lagrangian has a $SU(2)_L \times SU(2)_R$ chiral symmetry. The QCD vacuum spontaneously breaks this to the diagonal $SU(2)_V$ (isospin), with condensate:

$$\langle \Omega_{\text{QCD}} | \bar{q}q | \Omega_{\text{QCD}} \rangle = \langle \bar{u}u \rangle + \langle \bar{d}d \rangle \approx -(250 \text{ MeV})^3. \quad (34.85)$$

The chiral condensate is the Born output of the quark bilinear $\bar{\psi}\psi = \bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L$ in the QCD vacuum. It is non-zero because the QCD interaction mixes left- and right-handed quarks via gluon exchange, giving the vacuum a non-zero component in the mixed-chirality sector of the fermionic Fock space. The Goldstone bosons of the broken chiral symmetry are the pions (Section ??).

The gluon condensate. The QCD vacuum also has a non-zero gluon condensate:

$$\langle \Omega_{\text{QCD}} | \frac{\alpha_s}{\pi} G_{\mu\nu}^a G^{a\mu\nu} | \Omega_{\text{QCD}} \rangle \approx (330 \text{ MeV})^4, \quad (34.86)$$

arising from instantons (non-perturbative tunnelling events between different topological sectors of the gauge field) and from the non-perturbative structure of the QCD vacuum at scales

$\sim \Lambda_{\text{QCD}}$. The gluon condensate is the Born output of the chromoelectric and chromomagnetic field strengths in the QCD vacuum.

The QCD θ -term and CP violation. The QCD Lagrangian can include a topological term:

$$\mathcal{L}_\theta = \frac{\theta g_s^2}{32\pi^2} G_{\mu\nu}^a \tilde{G}^{a\mu\nu}, \quad (34.87)$$

where $\tilde{G}^{a\mu\nu} = \epsilon^{\mu\nu\rho\sigma} G_{\rho\sigma}^a/2$ is the dual field strength. This term is CP-odd: it violates both parity and time-reversal symmetry. The experimental bound on the neutron electric dipole moment gives $\theta < 10^{-10}$, which is a severe fine-tuning problem (the strong CP problem): why is θ so small? The Peccei-Quinn mechanism proposes a new axion field that dynamically drives $\theta \rightarrow 0$; the axion is then a pseudo-Goldstone boson with a mass $m_a \sim \Lambda_{\text{QCD}}^2/f_a$ where f_a is the Peccei-Quinn symmetry-breaking scale.

The QCD phase diagram. At high temperature $T > \Lambda_{\text{QCD}}$ or high baryon density, the QCD vacuum undergoes a phase transition to the quark-gluon plasma (QGP): a phase in which quarks and gluons are deconfined and the chiral condensate (34.85) melts ($\langle \bar{q}q \rangle \rightarrow 0$). The QGP was created in the universe a few microseconds after the Big Bang and is produced in heavy-ion collisions at RHIC and the LHC.

In the Fock-Born framework: the QCD phase transition is a change in the ground state of the QCD Hamiltonian. Below $T_c \approx 155 \text{ MeV}$ (the deconfinement temperature), the ground state $|\Omega_{\text{QCD}}\rangle$ has non-zero chiral condensate and physical states are colour-singlet hadrons. Above T_c , the ground state changes: the chiral condensate melts, and the physical states are (approximately) free quarks and gluons. The Hamiltonian H_{QCD} does not change; the vacuum does. This is the QCD analogue of the electroweak phase transition, except that the symmetry is restored at high temperature rather than broken.

Comparison: QED, electroweak, and QCD

Property	QED	Electroweak	QCD
Gauge group	$U(1)$	$SU(2)_L \times U(1)_Y$	$SU(3)_c$
Gauge bosons	γ (1, massless)	$W^\pm, Z^0$ (massive), γ (massless)	g^a (8, massless)
Self-interaction	None ($U(1)$ Abelian)	3-boson, 4-boson (broken $SU(2)$)	3-gluon, 4-gluon (unbroken $SU(3)$)
Higgs condensate	No	Yes ($v = 246$ GeV)	No
Symmetry breaking	None	$SU(2)_L \times U(1)_Y \rightarrow U(1)_{\text{EM}}$	$SU(2)_L \times SU(2)_R \rightarrow SU(2)_V$ (chiral)
β -function	> 0 (infrared free)	Mixed	< 0 (asymptotically free)
Coupling at m_Z	$\alpha \approx 1/129$	$\sin^2 \theta_W \approx 0.231$	$\alpha_s \approx 0.118$
Confinement	No	No	Yes (colour singlets only)
Physical d.o.f. per mode	$e^-: 2, e^+: 2, \gamma: 2$	$W^\pm: 2 \times 3, Z^0: 3, \gamma: 2, H: 1$, fermions	$q: 2 \times 3 \times 6, \bar{q}: \text{same}, g: 8 \times 2$
Bound states	Hydrogen (Coulomb)	WW, WZ (TeV?)	Hadrons (confinement)
Vacuum condensates	$\langle A_\mu \rangle = 0$	$\langle \phi^0 \rangle = v/\sqrt{2} \neq 0$	$\langle \bar{q}q \rangle \neq 0, \langle GG \rangle \neq 0$

All three theories are polynomials of creation and annihilation operators on their respective Fock spaces, with the same structural decomposition: $H = H_{\text{free}} + H_{\text{int}}$, where H_{free} is diagonal in the Fock basis (phase rotation) and H_{int} is off-diagonal (amplitude transfer between sectors). The differences are in the size of the Fock space (the number of field species and colour degrees of freedom), the self-interaction structure of the gauge bosons (Abelian versus non-Abelian), the vacuum structure (trivial versus condensate), and the confinement property (whether the physical Hilbert space is the full Fock space or the colour-singlet subspace).

Together, these three Hamiltonians on their three Fock spaces, connected by the Yukawa couplings to the Higgs condensate, are the Standard Model game engine. The Schrödinger equation $\dot{h}_t = -iH_{\text{SM}}h_t$ runs this engine; the Born layer $P(o_k) = |\langle o_k | h_t \rangle|^2$ renders the physical world we observe.

35 The Schrödinger Unrolled: Feynman Diagrams from the Dyson Series

We now have the complete Standard Model Fock space and Hamiltonian. The Fock hidden state h_t contains all particles and their quantum numbers; the Hamiltonian $H_{\text{SM}} = H_{\text{free}} + H_{\text{int}}$ is a polynomial of creation and annihilation operators. The Schrödinger equation $\dot{h}_t = -iH_{\text{SM}}h_t$ is the game engine. Before introducing the path integral, BRST, and the technical machinery of renormalization, we pause to ask the most basic question: what does the game engine actually *do* to the Fock hidden state?

The answer is the *Dyson series*: the Taylor expansion of $e^{-iH_{\text{SM}}t}$ in powers of H_{int} . Each term in this expansion is a specific sequence of elementary particle interactions, representable as a Feynman diagram. The Dyson series is the conceptually transparent face of quantum field theory — the one that makes the physics readable — and it precedes all the technical machinery that follows.

This section gives a symbolic treatment in the Fock operator framework. We are honest about the technical difficulties: the series is formal (UV divergences appear at loop order), the gauge orbit problem requires Faddeev-Popov, and operator ordering requires care. But the conceptual story is exact, and it is the most important thing to understand.

35.1 Splitting the Hamiltonian: free and interacting parts

Split H_{SM} into the free Hamiltonian and the interaction:

$$H_{\text{SM}} = H_{\text{free}} + H_{\text{int}}, \quad (35.1)$$

where:

$$H_{\text{free}} = \int \frac{d^3k}{(2\pi)^3} \sum_{\text{species}} \omega_{\mathbf{k}} a^\dagger(\mathbf{k}) a(\mathbf{k}) \quad (35.2)$$

sums over all particle species (quarks, leptons, gauge bosons, Higgs) with their respective dispersion relations $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m^2}$. The free Hamiltonian is diagonal in the Fock basis: it assigns an energy to each particle configuration and phase-rotates each basis vector at the corresponding rate. No particles are created or destroyed; the particle content of h_t is preserved.

The interaction Hamiltonian contains all the vertices of the Standard Model. Schematically, in the Fock operator language (suppressing coupling constants and spinor/colour/momentum indices for clarity):

$$\begin{aligned} H_{\text{int}} \supset & \underbrace{e b_e^\dagger b_e a_\gamma}_{\text{QED: } e^- \gamma \rightarrow e^-} + \underbrace{g_s b_q^\dagger b_q a_G}_{\text{QCD: } q G \rightarrow q} + \underbrace{g_s a_G^\dagger a_G a_G}_{\text{3-gluon}} + \underbrace{g_s^2 a_G^\dagger a_G^\dagger a_G a_G}_{\text{4-gluon}} \\ & + \underbrace{g b_\nu^\dagger b_\nu a_W^\dagger}_{\text{weak: } e^- \rightarrow \nu_e W^-} + \underbrace{y_f b_f^\dagger b_f a_h}_{\text{Yukawa: } f \bar{f} h} + \underbrace{\lambda a_h^\dagger a_h^\dagger a_h a_h}_{\text{Higgs self-coupling}} + \text{h.c.} + \dots \end{aligned} \quad (35.3)$$

Every term in (35.3) is a product of creation and annihilation operators describing a specific elementary process. Each term moves amplitude in the Fock hidden state from one basis vector to another: it destroys some particles and creates others, conserving energy, momentum, and all quantum numbers (colour, isospin, charge) required by the gauge symmetry.

35.2 The interaction picture: separating free and interacting dynamics

Working directly with $e^{-iH_{\text{SM}}t}$ is difficult because H_{free} and H_{int} do not commute. The *interaction picture* separates the two contributions cleanly. Define the interaction-picture state:

$$|\psi_I(t)\rangle = e^{iH_{\text{free}}t} |\psi(t)\rangle = e^{iH_{\text{free}}t} e^{-iH_{\text{SM}}t} |\psi(0)\rangle, \quad (35.4)$$

which satisfies:

$$\frac{d}{dt} |\psi_I(t)\rangle = -iH_{\text{int}}(t) |\psi_I(t)\rangle, \quad H_{\text{int}}(t) = e^{iH_{\text{free}}t} H_{\text{int}} e^{-iH_{\text{free}}t}. \quad (35.5)$$

In the interaction picture: the free Hamiltonian rotates the Fock basis vectors at their natural frequencies (the free propagation); the interaction Hamiltonian $H_{\text{int}}(t)$ drives transitions between basis vectors (the scattering). The state $|\psi_I(t)\rangle$ evolves only due to interactions, not due to free propagation.

The interaction-picture Hamiltonian $H_{\text{int}}(t)$ is the same polynomial (35.3) but with the creation and annihilation operators dressed by free-propagation phase factors:

$$a(\mathbf{k}, t) = e^{iH_{\text{free}}t} a(\mathbf{k}) e^{-iH_{\text{free}}t} = a(\mathbf{k}) e^{-i\omega_{\mathbf{k}}t}, \quad (35.6)$$

so the operators oscillate at their natural frequencies. This is the Heisenberg picture for free particles: each creation/annihilation operator carries the free phase $e^{\pm i\omega t}$.

35.3 The Dyson series: Taylor expanding the time evolution

The formal solution of (35.5) is the *time-ordered exponential*:

$$|\psi_I(t)\rangle = T \exp\left(-i \int_0^t dt' H_{\text{int}}(t')\right) |\psi_I(0)\rangle, \quad (35.7)$$

where T denotes time ordering: operators at earlier times appear to the right. Expanding (35.7) in powers of H_{int} :

$$\begin{aligned} |\psi_I(t)\rangle = & \left[1 - \underbrace{i \int_0^t dt_1 H_{\text{int}}(t_1)}_{\text{first order}} \right. \\ & \left. + \underbrace{\frac{(-i)^2}{2!} \int_0^t dt_1 \int_0^t dt_2 T[H_{\text{int}}(t_1)H_{\text{int}}(t_2)] + \cdots}_{\text{second order}} \right] |\psi_I(0)\rangle. \end{aligned} \quad (35.8)$$

This is the *Dyson series*. Each term is a specific power of H_{int} , integrated over the times at which the interactions occur. The n -th order term describes n successive interactions, each drawn from the vertex list (35.3).

What each term does to h_t . The zeroth-order term (1): the Fock hidden state is unchanged. No interactions occur. The initial particle configuration propagates freely to the final time.

The first-order term ($-i \int dt_1 H_{\text{int}}(t_1)$): one vertex from (35.3) is applied to h_t at time t_1 . For example, the QED vertex $e b_e^\dagger b_e a_\gamma$ applied to an initial state $|e^-(\mathbf{p}), \gamma(\mathbf{q})\rangle$ produces a state with one electron of momentum $\mathbf{p} + \mathbf{q}$ and no photon: the electron has absorbed the photon.

The second-order term: two vertices are applied at times $t_1 > t_2$. For example, two QED vertices applied to an initial electron state can describe the electron emitting a virtual photon (first vertex) and then reabsorbing it (second vertex): the electron self-energy. Or two vertices can describe an electron and a positron annihilating into a photon and then the photon creating a new electron-positron pair.

Each term in the Dyson series is a story about what happened to the particles between initial and final times. The full time-evolution operator $e^{-iH_{\text{SM}}t}$ is the sum of all such stories.

35.4 Feynman diagrams as terms in the Dyson series

A *Feynman diagram* is a graphical representation of one term in the Dyson series. The correspondence is exact:

External lines: the initial and final particles. The Fock basis vectors specifying the initial state $|\psi_I(0)\rangle$ and the final state we project onto.

Internal lines (propagators): a particle propagating freely between two interactions. Each internal line corresponds to a factor of the free propagator $\Delta(x - y) = \langle 0 | T[\hat{\phi}(x)\hat{\phi}(y)] | 0 \rangle$, which is the amplitude for a particle created at spacetime point y to propagate to point x . In the Fock operator language, this is the matrix element of $e^{-iH_{\text{free}}\tau}$ between two creation events.

Vertices: one application of H_{int} . Each vertex corresponds to one factor of $(-i\lambda)$ (or the appropriate coupling) from the vertex list (35.3), integrated over the spacetime point at which the interaction occurs.

The amplitude: the Born amplitude for going from initial to final state via the specific sequence of interactions described by the diagram. This is the matrix element:

$$\mathcal{M} = \langle \psi_f | (\text{specific term in Dyson series}) | \psi_i \rangle. \quad (35.9)$$

Sum over diagrams = sum over stories. The full scattering amplitude $\langle \psi_f | e^{-iHT} | \psi_i \rangle$ is the sum of (35.9) over all Feynman diagrams at all orders in the coupling. This sum is exactly the Dyson series (35.8). Each diagram is one interpretable story about what happened; the amplitude is the quantum superposition of all stories.

Examples from the Standard Model.

Electron-electron scattering (Møller scattering), $O(e^2)$: two QED vertices. An electron emits a photon; the photon is absorbed by the second electron. The photon propagator is the internal line; the two QED couplings e are the two vertices. The Fock hidden state, at second order in H_{int} , has amplitude moved from $|e^-(\mathbf{p}_1), e^-(\mathbf{p}_2)\rangle$ to $|e^-(\mathbf{p}'_1), e^-(\mathbf{p}'_2)\rangle$ via the virtual photon intermediate state.

Compton scattering, $O(e^2)$: two QED vertices. A photon is absorbed by an electron (first vertex); the electron emits a new photon (second vertex). Or equivalently: a photon is emitted first, then another photon is absorbed. Both orderings (both time orderings in the Dyson series) contribute; their sum is the Feynman diagram with two vertices connected by an internal electron propagator.

Quark-gluon scattering, $O(g_s)$: one QCD vertex. A quark absorbs a gluon and changes momentum. The colour indices are contracted by the $SU(3)$ generator T_{ij}^a .

Higgs production via gluon fusion, $O(y_t^2 g_s^2)$: two gluons fuse via a top quark loop and produce a Higgs boson. Four vertices: two QCD quark-gluon vertices and two Yukawa vertices. The top quark loop is a closed fermion line in the Fock operator language: a sequence of creation/annihilation events that returns to the initial particle species.

35.5 The Standard Model as a story generator

The Dyson series makes the Standard Model's dynamics completely transparent. The Fock hidden state h_t is a superposition of all particle configurations. The Hamiltonian H_{int} is a

vocabulary of elementary interaction verbs: “electron absorbs photon”, “quark emits gluon”, “top quark couples to Higgs”, “W boson mediates beta decay”. The Dyson series is the story generator: it produces, at each order in the coupling, all possible sequences of these verbs, each sequence describing a specific physical process. The Born rendering layer measures the amplitude of each story and produces a classical outcome.

This is quantum field theory at its most interpretable:

- Each Fock basis vector is a monosemantic world configuration (exactly these particles with exactly these quantum numbers).
- Each vertex in H_{int} is a monosemantic elementary process (one specific type of interaction).
- Each Feynman diagram is a monosemantic story (one specific sequence of interactions).
- The full amplitude is a polysemantic superposition (the quantum sum over all stories).
- The Born rendering layer extracts one classical story from the superposition.

The transformer generates text by processing a polysemantic residual stream with dense weight matrices whose elementary operations have no pre-assigned meaning. The Standard Model generates physical events by processing a polysemantic Fock hidden state with a sparse, structured Hamiltonian whose elementary operations are named vertices with fixed physical interpretations. Both produce rich, structured outputs from simple iterative rules. The Standard Model’s outputs are more interpretable because its rules were written in interpretable form from the beginning.

35.6 What the diagrams are: Fock-layer bookkeeping

The language of the previous subsection was deliberately loose: “the electron emits a photon,” “the gluon splits into a quark pair,” “two gluons fuse to produce a Higgs.” This language is seductive because it makes the Dyson series narrative and memorable. But it is a conflation, and we must now correct it.

The conflation. Every phrase of the form “particle A emits particle B ” or “particles A and B annihilate into C ” narrates a Fock-layer computation in Born-layer language. The Dyson series is a Taylor expansion of a matrix element of e^{-iHT} : each term is a specific product of creation and annihilation operators, integrated over intermediate times, contracted with propagators. The “emission” is an annihilation operator acting on one Fock basis vector and a creation operator acting on another. The “virtual photon” is a propagator $G_0^{(2)}(x - y)$ —a Green’s function, a mathematical factor in the expansion, not a Born-layer event.

No observer detects the “emission.” No Born rendering occurs at the vertex. No Bohr trimming resets the state. The entire Dyson series is one computation of one matrix element of the time-evolution operator. The Fock hidden state is one vector, rotating under $e^{-iH_{\text{SM}}T}$, and we are decomposing that rotation into a sum of elementary algebraic operations.

The correct language. Each term in the Dyson series is a *contribution to a matrix element*: a specific product of coupling constants, propagators, and vertex factors that, when summed over all terms and squared, gives the Born probability $P(i \rightarrow f) = |\langle f | S | i \rangle|^2$.

A Feynman diagram is a graphical notation for one such term. The vertices are coupling constants; the internal lines are propagators; the external lines are the initial and final states. The diagram is a bookkeeping device for the algebra of the Dyson series, exactly as a circuit diagram is a bookkeeping device for the algebra of Kirchhoff’s laws. The circuit diagram does not depict electrons flowing; the Feynman diagram does not depict particles interacting.

Why the conflation persists. Narrating diagrams as stories is pedagogically useful because

it makes the algebra memorable: “the electron emits a photon” is easier to remember than “the first-order Dyson term for the QED vertex applied to an electron Fock state produces a two-particle state via the coupling $e b_e^\dagger b_e a_\gamma$.” The narrative is a mnemonic, not a description.

The conflation becomes harmful when it generates questions that the formalism cannot answer: “How does the electron decide to emit the photon?” “What happens to the virtual photon between the two vertices?” “How long does the virtual pair exist?” These questions have no answers because the premises are wrong: no emission occurs, no virtual photon exists as a physical entity, and no pair is created. The diagram is one term in a sum; the sum is one matrix element; the matrix element squared is the Born probability. Everything between the initial state and the final probability is Fock-layer bookkeeping.

35.7 The complete rendering pipeline

We now state the complete pipeline from the Schrödinger equation to the detector click, labelling each stage by its layer.

Stage 0: the engine (Fock layer). The Fock hidden state h_t evolves under $H_{\text{SM}}(a)$: $\dot{h}_t = -iH_{\text{SM}}h_t$. The Hamiltonian is a polynomial of creation and annihilation operators on $\mathcal{F}_a^{\text{SM,phys}}$, with bare parameters $\theta(a)$. The evolution is deterministic, unitary, and linear. No events occur in the Fock layer.

Stage 1: the Dyson series (Fock layer). Taylor-expand $e^{-iH_{\text{SM}}T}$ in powers of H_{int} to obtain the S -matrix: $S = T \exp\left(-i \int H_{\text{int}}^I dt\right)$. Each order n contributes a sum of n -vertex diagrams. The S -matrix element $\langle f | S | i \rangle$ is a complex number: the Fock-layer amplitude for transition from $|i\rangle$ to $|f\rangle$.

Stage 2: the amplitude $\mathcal{M}$ (Fock layer). The connected, amputated S -matrix element $\mathcal{M} = \langle f | iT | i \rangle$ (where $S = 1 + iT$) is extracted from $\langle f | S | i \rangle$ by removing the trivial forward-scattering contribution and the external propagators (LSZ reduction). $\mathcal{M}$ is a complex function of the momenta and quantum numbers of the initial and final particles. It contains phase information. It can interfere. It is not observed.

Stage 3: the Born rule (Born layer). The Born probability:

$$P(i \rightarrow f) = |\mathcal{M}(i \rightarrow f)|^2 \times (\text{phase space}). \quad (35.10)$$

This is the squared modulus of the Fock-layer amplitude, integrated over the final-state kinematics. The phase information is erased. The result is a non-negative real number: the differential cross section $d\sigma/d\Omega$.

Stage 4: the detector (Born layer). The detector is a macroscopic system with $\sim 10^{23}$ degrees of freedom. When a particle deposits energy in the detector, a macroscopic observable (current, light, heat) transitions to a new eigenvalue. This eigenvalue is rendered by the Born rule with enormous degeneracy ($\sim e^{S/k_B}$ microstates per macroscopic reading). The detector click is a Born-layer event.

Stage 5: the histogram (Born layer). The experimentalist collects $\sim 10^6$ detector events, bins them by kinematic variables ($\sqrt{s}$, θ , p_T , invariant mass), and produces a histogram. The histogram is a Born-layer frequency distribution: an estimate of the Born probability $P(i \rightarrow f)$ from finite samples. The histogram is compared to the SM prediction $d\sigma/d\Omega$ from Stages 1–3.

The pipeline diagram:

	Stage	Layer	Object
0	Schrödinger equation	Fock	$\dot{h}_t = -iHh_t$ (one vector, rotating)
1	Dyson series	Fock	$S = T \exp(-i \int H_I)$ (sum over diagrams)
2	Amplitude	Fock	$\mathcal{M} \in \mathbb{C}$ (complex, interferes)
3	Born rule	$\longrightarrow$	$ \mathcal{M} ^2$ (real, positive, no phase)
4	Detector click	Born	Macroscopic eigenvalue rendered
5	Histogram	Born	Frequency estimate of $P(i \rightarrow f)$

The arrow at Stage 3 is the one-way valve: it separates the Fock layer (Stages 0–2) from the Born layer (Stages 4–5). Everything above the arrow is internal to the engine; everything below is rendered to the observer. The squared modulus $|\cdot|^2$ is the gate.

No information flows backward. The observer cannot infer $\mathcal{M}$ from $|\mathcal{M}|^2$: the phase is lost. The detector cannot probe the intermediate states of the Dyson series: no Born rendering occurs at the vertices. The Fock layer is opaque to the Born layer. The only information that crosses the gate is the modulus squared.

This is why “which path did the virtual photon take?” and “how long did the virtual pair exist?” are unanswerable: they ask about Fock-layer internals that the Born-layer gate does not transmit. The gate transmits $|\mathcal{M}|^2$ and nothing else.

35.8 The monosemantic Hamiltonian versus the polysemantic residual stream

The analogy between the Standard Model and a recurrent neural network, which runs throughout this book, reaches its most precise form in the Dyson series. The comparison is worth making explicit.

The Standard Model RNN. The hidden state $h_t \in \mathcal{F}_a^{\text{SM}}$ is a superposition of Fock basis vectors, each of which is a specific particle configuration with definite quantum numbers. The Fock basis is *monosemantic*: each basis vector has a fixed, pre-assigned physical meaning (“two up quarks, one down quark, one gluon, all with specified momenta, spins, and colours”). The Hamiltonian H_{SM} is a sparse polynomial whose terms are named vertices with fixed physical interpretations (“QED coupling,” “strong vertex,” “Yukawa”). The Dyson series is a sum over interpretable stories, each a sequence of named vertices.

A large language model. The residual stream $h_t \in \mathbb{R}^d$ of a transformer is a dense, high-dimensional vector. The basis is not monosemantic: the individual coordinates have no pre-assigned meaning. The weight matrices W_Q, W_K, W_V, W_O are dense matrices whose entries have no physical interpretation. The forward pass is a sequence of matrix multiplications and nonlinearities, none of which has a named meaning.

The comparison.

	Standard Model	Transformer
Hidden state	$h_t \in \mathcal{F}_a^{\text{SM}}$	$h_t \in \mathbb{R}^d$
Basis	Monosemantic (particle configs)	Polysemantic (no named meaning)
Weight matrix	H_{SM} : sparse, named vertices	W : dense, unnamed entries
“Forward pass”	Dyson series: sum over named stories	Matrix multiplications: no named steps
Output layer	Born rule: $ \mathcal{M} ^2$	Softmax: $p(\text{token})$
Interpretability	High: each diagram is a story	Low: each forward pass is opaque

The Standard Model is the most interpretable neural network in existence: its hidden state has a named basis, its weight matrix has named entries, and its forward pass decomposes into a sum of named stories (the Dyson series). The transformer is the least interpretable: its hidden state has no named basis, its weight matrix has no named entries, and its forward pass does not decompose into named steps.

The “mechanistic interpretability” programme in AI research aims to find the monosemantic features hidden inside polysemantic residual streams—to decompose the transformer’s dense representations into sparse, named, physically meaningful components. The Standard Model is the existence proof that this decomposition is possible: the Fock basis and the vertex structure of H_{SM} are the monosemantic features of the quantum field theory. Nature’s neural network was born interpretable.

35.9 Technical difficulties: what the Dyson series cannot yet do

The Dyson series gives the correct conceptual picture, but it faces several technical difficulties that require the machinery developed in subsequent sections. We list them honestly, so the reader knows what is coming and why.

UV divergences. At second order and beyond, the Dyson series generates integrals over the momenta of virtual (internal) particles. These integrals diverge at high momentum: the loop integral $\int d^4k/(k^2 - m^2)^2$ diverges logarithmically. This is not a failure of the conceptual story; it is a failure of the naive integration. Regularization (dimensional regularization) renders the integrals finite, and renormalization absorbs the divergences into the parameters of the Hamiltonian. The conceptual story (electron emits and reabsorbs a virtual photon) is correct; the technical difficulty is computing the amplitude of that story without encountering infinities.

Gauge orbit overcounting. The gauge fields $A_\mu(x)$ in the Hamiltonian (35.3) are not physical observables: configurations related by gauge transformations are physically identical. The Dyson series, written in terms of A_μ , integrates over all gauge configurations including multiple copies of each physical configuration. This overcounting generates infinities that have nothing to do with physics. The Faddeev-Popov procedure removes the overcounting by selecting one representative per gauge orbit, at the cost of introducing ghost fields. In the Fock operator language, the physical Hilbert space is the BRST cohomology $\ker Q/\text{im } Q$: only gauge-invariant combinations of creation operators create physical states. A symbolic treatment that ignores this problem is exact at the level of gauge-invariant observables (such as S -matrix elements between colour-singlet states) but requires the full Faddeev-Popov machinery to be rigorous.

Operator ordering. When H_{int} contains products of fields at the same spacetime point, the ordering of creation and annihilation operators matters. Normal ordering (all $a^\dagger$ to the left of all a) is the standard convention, but it introduces finite corrections (the normal-ordering constants)

that must be tracked carefully. The path integral formulation avoids explicit operator ordering by working with classical fields, at the cost of reintroducing ordering issues through the measure $\mathcal{D}\phi$.

Non-perturbative effects. The Dyson series is a perturbative expansion in the coupling constants. At strong coupling (e.g. QCD at low energies), the series diverges and perturbation theory breaks down. Non-perturbative effects — confinement, the QCD condensate, instantons — require methods beyond the Dyson series: the lattice path integral, the large- N expansion, or effective field theories.

What the symbolic treatment gives us. Despite these difficulties, the Dyson series at the symbolic level gives the complete conceptual picture:

1. The Fock hidden state is a superposition of particle configurations.
2. The Hamiltonian is a polynomial of creation and annihilation operators, with each term describing a specific elementary interaction.
3. The Schrödinger equation unrolls the Dyson series: at each order in the coupling, specific sequences of elementary interactions occur, described by Feynman diagrams.
4. The Born rendering layer measures the amplitude of the final Fock configuration and produces a classical outcome.
5. The path integral is the generating functional for the Dyson series: it produces the same Feynman diagrams by a different computational route, better suited for handling gauge orbits and UV divergences systematically.

The technical sections that follow — path integral, Faddeev-Popov, BRST, renormalization — are solutions to the technical difficulties listed above. They do not change the conceptual story; they make it rigorous.

Dyson series term	Feynman diagram	Physical story
$(-i)^0 \mathbf{1}$	No vertices	Free propagation; no interactions
$(-i)^1 \int dt_1 H_{\text{int}}(t_1)$	One vertex	One elementary interaction
$(-i)^2 \int dt_1 dt_2 T[H_{\text{int}}^2]$	Two vertices	Two successive interactions
$(-i)^n \int \dots$	n vertices	n successive interactions
Sum over all orders	Sum over all diagrams	Quantum superposition of all stories
UV divergence at loop order	Loop diagram	Virtual particle momentum unbounded
Gauge orbit overcounting	Unphysical diagrams	Faddeev-Popov needed
Non-perturbative: e^{-1/g^2}	Not visible in any diagram	Instanton, confinement

36 Path Integral for the Standard Model

In the scalar field section, we derived the path integral by slicing the time evolution operator e^{-iHT} into N thin steps, inserting a resolution of the identity at each slice, and taking $N \rightarrow \infty$. The result was an integral over field configurations $\phi(x, t)$, weighted by $e^{iS[\phi]}$ where $S = \int d^4x \mathcal{L}$ is the action. We emphasised that the integration variables $\phi(x, t)$ at each spacetime point are not classical field values but *eigenvalues of field eigenstates* used to resolve the identity in Fock space.

The Standard Model path integral follows exactly the same logic. The only new ingredient is that the fields are no longer scalar: they carry spinor, colour, and isospin indices, and the fermion fields require a new kind of number — *Grassmann numbers* — to resolve the identity in the fermionic Fock space. We build the path integral species by species, starting with the Dirac fermion.

36.1 Recap: the scalar path integral as a matrix product

For a free scalar field, the quantum-mechanical amplitude for evolving from state $|\phi_i\rangle$ to state $|\phi_f\rangle$ in time T is:

$$\langle \phi_f | e^{-iHT} | \phi_i \rangle = \int_{\phi(0)=\phi_i}^{\phi(T)=\phi_f} \mathcal{D}\phi e^{iS[\phi]}, \quad (36.1)$$

derived by slicing T into N steps of width $\varepsilon = T/N$ and inserting at each step the identity:

$$\mathbf{1} = \int d\phi_k |\phi_k\rangle \langle \phi_k| \quad (36.2)$$

where $|\phi_k\rangle$ is the eigenstate of the field operator $\hat{\phi}(\mathbf{x})$ with eigenvalue $\phi_k(\mathbf{x})$. Three points are essential:

1. The integration variable $\phi_k(\mathbf{x})$ at each time slice t_k is the *eigenvalue* of the field operator $\hat{\phi}(\mathbf{x})$ in the eigenstate $|\phi_k\rangle$. It is a real number at each spatial point $\mathbf{x}$, but it is not a classical field — it is a label for a Fock-space eigenstate, introduced purely to resolve the identity.
2. The matrix element $\langle \phi_{k+1} | e^{-i\mathcal{H}\varepsilon} | \phi_k \rangle$ between adjacent slices evaluates to $e^{i\varepsilon \mathcal{L}(\phi_k, \partial_\mu \phi_k)}$ in the limit $\varepsilon \rightarrow 0$, which is the link between the Hamiltonian (matrix) and the Lagrangian (action) pictures.
3. The product of N such matrix elements, integrated over all intermediate field configurations, is the path integral (36.1). The path integral is a continuous matrix product: the limit of an N -fold matrix product as $N \rightarrow \infty$.

Every element of this structure reappears for gauge fields and fermions. The new ingredient is the resolution of the identity in the fermionic Fock space, which requires Grassmann numbers.

36.2 Fermionic Fock space and its eigenstates

For the scalar field, the field operator $\hat{\phi}(\mathbf{x})$ has eigenstates $|\phi\rangle$ with real eigenvalues $\phi(\mathbf{x})$. These eigenstates are coherent states in the bosonic Fock space, and their existence follows from the commutation relations $[\hat{\phi}(\mathbf{x}), \hat{\phi}(\mathbf{y})] = 0$.

For a fermionic field, the situation is fundamentally different. The Dirac field operator $\hat{\Psi}(\mathbf{x})$ satisfies *anticommutation* relations:

$$\{\hat{\Psi}^A(\mathbf{x}), \hat{\Psi}^{B\dagger}(\mathbf{y})\} = \delta^{AB} \delta^3(\mathbf{x} - \mathbf{y}), \quad \{\hat{\Psi}^A(\mathbf{x}), \hat{\Psi}^B(\mathbf{y})\} = 0. \quad (36.3)$$

These anticommutation relations mean that $\hat{\Psi}^A(\mathbf{x})^2 = 0$: the field operator squares to zero. This is the algebraic expression of Pauli exclusion: at most one fermion per mode.

Why ordinary numbers cannot label fermionic eigenstates. For the bosonic field, we asked: is there a state $|\psi\rangle$ such that $\hat{\phi}(\mathbf{x})|\psi\rangle = \phi(\mathbf{x})|\psi\rangle$ for some ordinary (commuting) number $\phi(\mathbf{x})$? The answer is yes: bosonic coherent states.

For the fermionic field, we ask: is there a state $|\eta\rangle$ such that $\hat{\Psi}(\mathbf{x})|\eta\rangle = \eta(\mathbf{x})|\eta\rangle$ for some number $\eta(\mathbf{x})$? If such a number exists, it must satisfy the same algebra as the operator. In particular, squaring both sides of the eigenvalue equation:

$$0 = \hat{\Psi}(\mathbf{x})^2 |\eta\rangle = \eta(\mathbf{x})^2 |\eta\rangle, \quad (36.4)$$

so $\eta(\mathbf{x})^2 = 0$. An ordinary real or complex number with this property does not exist (only $\eta = 0$ satisfies it for ordinary numbers). The eigenvalues of fermionic field operators cannot be ordinary numbers. They must be *Grassmann numbers*.

36.3 Grassmann numbers: algebra and calculus

Grassmann numbers are the mathematical objects that extend the concept of number to include anticommuting quantities. They were introduced by Hermann Grassmann in 1844 as an abstract algebraic construction; they became indispensable in physics only when it was recognized that they are exactly the right language for fermionic path integrals.

Definition: generators of a Grassmann algebra

A *Grassmann algebra* over $\mathbb{C}$ with n generators $\theta_1, \theta_2, \dots, \theta_n$ is the algebra generated by these elements subject to the single rule:

$$\theta_i \theta_j + \theta_j \theta_i = 0 \quad \text{for all } i, j. \quad (36.5)$$

Setting $i = j$: $\theta_i^2 + \theta_i^2 = 2\theta_i^2 = 0$, so $\theta_i^2 = 0$. Every generator squares to zero.

A general element of the Grassmann algebra is a polynomial in the generators. Because $\theta_i^2 = 0$, the polynomial terminates: each generator appears at most once in any monomial. For a single generator θ , the most general element is:

$$f(\theta) = a + b\theta, \quad a, b \in \mathbb{C}. \quad (36.6)$$

There is no θ^2 term because $\theta^2 = 0$. For two generators θ_1, θ_2 :

$$f(\theta_1, \theta_2) = a + b_1\theta_1 + b_2\theta_2 + c\theta_1\theta_2, \quad a, b_1, b_2, c \in \mathbb{C}. \quad (36.7)$$

For n generators, the algebra has dimension 2^n as a vector space over $\mathbb{C}$.

Grassmann numbers are not numbers in the ordinary sense. They have no ordering ($\theta > 0$ is meaningless), no magnitude, and no real-number representation. They are purely algebraic entities defined by their multiplication rules. What makes them useful is that these rules perfectly mirror the anticommutation relations (36.3) of fermionic field operators.

Even and odd elements

Elements of the Grassmann algebra split into *even* (bosonic) and *odd* (fermionic) parts:

- *Even* elements contain an even number of generators in each monomial: $a, \theta_i\theta_j, \theta_i\theta_j\theta_k\theta_l, \dots$ These *commute* with everything in the algebra.

- *Odd* elements contain an odd number of generators: $\theta_i, \theta_i\theta_j\theta_k, \dots$. These *anticommute* with other odd elements: $\theta_i\theta_j = -\theta_j\theta_i$.

A single Grassmann generator θ is an odd element. The product $\theta_1\theta_2$ is an even element. This even/odd grading is the algebraic shadow of the boson/fermion distinction in physics.

Grassmann calculus: differentiation

Differentiation with respect to a Grassmann variable is defined by:

$$\frac{\partial}{\partial\theta}(a + b\theta) = b, \quad \frac{\partial}{\partial\theta}\theta = 1, \quad \frac{\partial}{\partial\theta}1 = 0. \quad (36.8)$$

For multiple generators, differentiation anticommutes with generators to its right:

$$\frac{\partial}{\partial\theta_i}(\theta_j\theta_k) = \delta_{ij}\theta_k - \delta_{ik}\theta_j. \quad (36.9)$$

More precisely: $\partial/\partial\theta_i$ anticommutes with all Grassmann generators, so it “passes through” odd elements picking up a minus sign:

$$\frac{\partial}{\partial\theta_i}(\theta_j f) = \delta_{ij}f - \theta_j \frac{\partial f}{\partial\theta_i}. \quad (36.10)$$

This is the graded Leibniz rule.

Grassmann calculus: integration (Berezin integral)

Integration over Grassmann variables was defined by Felix Berezin in 1966. The definition looks paradoxical at first: it is defined to equal differentiation. For a single Grassmann variable θ :

$$\int d\theta \, 1 = 0, \quad \int d\theta \, \theta = 1. \quad (36.11)$$

For a function $f(\theta) = a + b\theta$:

$$\int d\theta \, f(\theta) = \int d\theta \, (a + b\theta) = b. \quad (36.12)$$

This is identical to $\partial f/\partial\theta = b$. Grassmann integration is Grassmann differentiation.

Why this definition? The Berezin integral is not defined to generalise the Riemann integral — it is defined to satisfy two properties that are needed for the path integral:

1. **Linearity:** $\int d\theta (af + bg) = a \int d\theta f + b \int d\theta g$.
2. **Translation invariance:** For any Grassmann constant η , $\int d\theta f(\theta + \eta) = \int d\theta f(\theta)$. (This is the Grassmann analogue of $\int_{-\infty}^{\infty} f(x + a)dx = \int_{-\infty}^{\infty} f(x)dx$ for ordinary integrals.)

Translation invariance forces $\int d\theta \, 1 = 0$: if $f(\theta) = 1$ (constant), then $f(\theta + \eta) = 1$ for any η , so translation invariance requires $\int d\theta \, 1 = \int d\theta \, 1$, which is trivially true — but the only linear functional satisfying this for all η is the zero functional (since for $f(\theta) = 1$, shifting by η gives $f(\theta + \eta) = 1$ and the integral must not change when η is arbitrary). More carefully: $f(\theta + \eta) = 1$ but also $f(\theta + \eta) = a + b(\theta + \eta) = a + b\theta + b\eta$; for $f = 1$, $a = 1$, $b = 0$, and the integral of any constant must be zero for translation invariance to hold non-trivially.

For two conjugate Grassmann variables $\bar{\theta}, \theta$:

$$\int d\bar{\theta} d\theta \, e^{-\bar{\theta}M\theta} = M \quad (36.13)$$

for a complex number M . Compare this to the bosonic Gaussian:

$$\int \frac{d\bar{z} dz}{2\pi i} e^{-\bar{z} M z} = \frac{1}{M}. \quad (36.14)$$

The Grassmann result is the *inverse* of the bosonic result. This inversion is the algebraic source of the difference between bosonic and fermionic determinants in the path integral.

For n pairs of conjugate Grassmann variables $(\bar{\theta}_i, \theta_j)$ and an $n \times n$ complex matrix M :

$$\int \prod_{i=1}^n d\bar{\theta}_i d\theta_i e^{-\bar{\theta}_i M^i_j \theta^j} = \det M. \quad (36.15)$$

Again compare to the bosonic result: $\int \prod_i d\bar{z}_i dz_i e^{-\bar{z}_i M^i_j z^j} = (2\pi)^n / \det M$. Grassmann integration gives $\det M$ (in the numerator); bosonic integration gives $1/\det M$ (in the denominator). This is why fermion loops contribute a factor of $\det M$ to the path integral, while boson loops contribute $1/\det M$.

Fermionic coherent states: resolving the identity

A *fermionic coherent state* for a single fermionic mode with creation operator $b^\dagger$ and annihilation operator b is:

$$|\eta\rangle = e^{-\eta b^\dagger} |0\rangle = |0\rangle - \eta |1\rangle, \quad (36.16)$$

where η is a Grassmann number and $|0\rangle, |1\rangle$ are the zero- and one-particle Fock states. The state $|\eta\rangle$ is an eigenstate of the annihilation operator b :

$$b |\eta\rangle = \eta |\eta\rangle. \quad (36.17)$$

Proof. $b |\eta\rangle = b(|0\rangle - \eta |1\rangle) = 0 - \eta b |1\rangle = -\eta |0\rangle$. But also $\eta |\eta\rangle = \eta |0\rangle - \eta^2 |1\rangle = \eta |0\rangle$ (since $\eta^2 = 0$). So $b |\eta\rangle = -\eta |0\rangle$ while $\eta |\eta\rangle = \eta |0\rangle$. These agree only if we adopt the sign convention that Grassmann numbers anticommute with fermionic operators, i.e. $b \eta = -\eta b$. With this convention, the eigenvalue equation (36.17) holds. $\square$

The eigenvalue of b on $|\eta\rangle$ is the Grassmann number η , not an ordinary complex number. This is the precise statement that *fermionic field eigenvalues are Grassmann numbers*: the eigenvalues of $\hat{\Psi}(\mathbf{x})$ on fermionic coherent states are Grassmann-valued functions of $\mathbf{x}$.

The resolution of the identity in the fermionic Fock space is:

$$\mathbf{1} = \int d\bar{\eta} d\eta e^{-\bar{\eta} \eta} |\eta\rangle \langle \eta|, \quad (36.18)$$

where $\bar{\eta}$ is the independent conjugate Grassmann variable (corresponding to $b^\dagger$). This is the fermionic analogue of the bosonic resolution $\mathbf{1} = \int d\phi |\phi\rangle \langle \phi|$.

The path integral integration variable is the coherent state eigenvalue. Just as the bosonic path integral integrates over $\phi(\mathbf{x}, t)$ (the eigenvalue of $\hat{\phi}(\mathbf{x})$ in the bosonic coherent state $|\phi\rangle$), the fermionic path integral integrates over $\psi(\mathbf{x}, t)$ (the Grassmann eigenvalue of $\hat{\Psi}(\mathbf{x})$ in the fermionic coherent state $|\eta\rangle$). In both cases, the integration variable is not a classical field: it is the eigenvalue of a Fock-space eigenstate used to resolve the identity.

Remark. [ML analogy: Grassmann numbers as anticommuting feature dimensions]

In a neural network processing fermionic data, one might want a feature space where combining the same feature twice gives zero. Grassmann numbers implement exactly this: $\theta^2 = 0$ means

the same Grassmann variable cannot appear twice in any product. This is the algebraic implementation of Pauli exclusion at the level of the field variables. The Berezin integral — which equals differentiation — is then the natural dual operation: extracting the coefficient of θ in a Grassmann polynomial is both differentiation and integration. The symmetry between differentiation and integration, which is exceptional in ordinary calculus, is the rule in Grassmann calculus.

36.4 The Dirac path integral

With fermionic coherent states and the resolution of the identity (36.18) in hand, the Dirac path integral follows by the same time-slicing procedure as the scalar case.

Time slicing for fermions

The Hamiltonian for a free Dirac field is:

$$H_{\text{Dirac}} = \int d^3\mathbf{x} \, \hat{\Psi}^A(\mathbf{x}) (-i\gamma^i \partial_i + m)^A_B \hat{\Psi}^B(\mathbf{x}), \quad (36.19)$$

where $A, B \in \{1, 2, 3, 4\}$ are Dirac indices and $i \in \{1, 2, 3\}$ are spatial indices. The amplitude for evolving from fermionic Fock state $|\eta_i\rangle$ to $|\eta_f\rangle$ in time T is:

$$\langle \eta_f | e^{-iH_{\text{Dirac}}T} | \eta_i \rangle. \quad (36.20)$$

Slice T into N steps of width $\varepsilon = T/N$. Insert the fermionic resolution of the identity (36.18) at each time slice $t_k = k\varepsilon$:

$$\langle \eta_f | e^{-iH_{\text{Dirac}}T} | \eta_i \rangle = \int \prod_{k=1}^{N-1} (d\bar{\eta}_k d\eta_k e^{-\bar{\eta}_k \eta_k}) \prod_{k=0}^{N-1} \langle \eta_{k+1} | e^{-iH_{\text{Dirac}}\varepsilon} | \eta_k \rangle. \quad (36.21)$$

Each factor $\langle \eta_{k+1} | e^{-iH\varepsilon} | \eta_k \rangle$ is a matrix element between adjacent coherent states. In the limit $\varepsilon \rightarrow 0$, using $e^{-iH\varepsilon} \approx 1 - iH\varepsilon$ and evaluating the coherent state matrix element:

$$\langle \eta_{k+1} | e^{-iH_{\text{Dirac}}\varepsilon} | \eta_k \rangle \approx e^{\bar{\eta}_{k+1} \eta_k} \exp \left[i\varepsilon \bar{\eta}_{k+1}^A (i\partial_t - (-i\gamma^i \partial_i + m))^A_B \eta_k^B \right]. \quad (36.22)$$

The term $\bar{\eta}_{k+1} \eta_k$ in the exponent comes from the overlap $\langle \eta_{k+1} | \eta_k \rangle = e^{\bar{\eta}_{k+1} \eta_k}$ of adjacent coherent states, which is the Grassmann analogue of the scalar coherent state overlap. Taking the continuum limit $\varepsilon \rightarrow 0$, $N \rightarrow \infty$ with $N\varepsilon = T$ fixed:

$$\langle \eta_f | e^{-iH_{\text{Dirac}}T} | \eta_i \rangle = \int \mathcal{D}\bar{\psi} \mathcal{D}\psi \exp \left[i \int_0^T d^4x \, \bar{\psi}^A(x) (i\gamma^\mu \partial_\mu - m)^A_B \psi^B(x) \right], \quad (36.23)$$

where $\mathcal{D}\bar{\psi} \mathcal{D}\psi = \prod_{x,A} d\bar{\psi}^A(x) d\psi^A(x)$ is the Grassmann functional measure — a product of Berezin integrals over all spacetime points and all Dirac indices.

The exponent is $iS_{\text{Dirac}}[\bar{\psi}, \psi]$ where:

$$S_{\text{Dirac}}[\bar{\psi}, \psi] = \int d^4x \, \bar{\psi}^A(x) (i\gamma^\mu \partial_\mu - m)^A_B \psi^B(x). \quad (36.24)$$

What the Dirac path integral variables are. The fields $\psi^A(x)$ and $\bar{\psi}^A(x)$ in the path integral are Grassmann-valued functions of spacetime: at each point x and each Dirac index A , $\psi^A(x)$ is a Grassmann number, not a complex number. They are the *eigenvalues* of the Dirac field operator $\hat{\Psi}^A(x)$ on fermionic coherent states, introduced to resolve the identity at each

time slice. They are not classical fields; the Grassmann algebra has no classical limit ($\theta^2 = 0$ has no analogue in ordinary numbers).

Evaluating the free Dirac path integral. The action (36.24) is quadratic in the Grassmann fields. Using the Grassmann Gaussian formula (36.15):

$$\int \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{i \int d^4x \bar{\psi}(i\cancel{\partial} - m)\psi} = \det(i\cancel{\partial} - m), \quad (36.25)$$

where $\det(i\cancel{\partial} - m)$ is the functional determinant of the Dirac operator. For bosons the result would be $[\det(i\cancel{\partial} - m)]^{-1}$; for fermions it is the determinant itself (no inverse), because Grassmann Gaussians give $\det M$ rather than $1/\det M$. This functional determinant encodes all the quantum effects of the free Dirac field.

36.5 The round trip: Lagrangian $\rightarrow$ Hamiltonian $\rightarrow$ time slicing $\rightarrow$ Lagrangian

We have now seen the Dirac path integral derived from the Hamiltonian by time slicing. It is worth pausing to examine the round trip carefully, because two things happen along the way that are easy to miss and important to understand: the field variables change, and the Lagrangian that re-emerges is not the same object as the one we started with.

The starting point: the Hamiltonian on Fock space. We begin with the Hamiltonian $H[\hat{\psi}, \hat{\pi}]$ — a polynomial of the field operator $\hat{\psi}$ and its conjugate momentum $\hat{\pi}$, acting on the Fock space $\mathcal{F}$. For the Dirac field, the Hamiltonian is:

$$H = \int d^3\mathbf{x} \hat{\bar{\psi}}(-i\vec{\gamma} \cdot \nabla + m)\hat{\psi}, \quad (36.26)$$

a specific polynomial of the fermionic creation and annihilation operators $b^\dagger(\mathbf{k}, s)$ and $b(\mathbf{k}, s)$ (and their antiparticle counterparts). This is the game engine: the Schrödinger equation $\dot{h}_t = -iHh_t$ rotates the Fock hidden state under (36.26), one Planck step at a time. All physics is in H ; the Lagrangian has not appeared yet.

Time slicing: inserting resolutions of the identity. The transition amplitude $\langle f | e^{-iHT} | i \rangle$ is computed by splitting $e^{-iHT} = (e^{-iH\varepsilon})^N$ and inserting a resolution of the identity between each factor. For a bosonic scalar field, the resolution of the identity in the coherent state basis:

$$\mathbf{1} = \int \mathcal{D}\phi | \phi \rangle \langle \phi | \quad (36.27)$$

introduces the integration variable ϕ at each time slice: an ordinary complex number (for a complex scalar) or real number (for a real scalar) at each spacetime point. This is the field variable of the path integral — and it is crucial to understand what it is. It is *not* a classical field configuration that the quantum field “takes on” at time t_k . It is an integration variable: the label of the coherent state basis vector $|\phi_{t_k}\rangle$ inserted at time step k to decompose the identity operator. The path $(i, \phi_1, \dots, \phi_{N-1}, f)$ is a sequence of integration labels, one per time slice, not a sequence of physical field configurations that the system passes through. The integral over these labels is a matrix multiplication, not a sum over histories.

The Gaussian integral over the conjugate momentum. For a scalar field with Hamiltonian $H = \int d^3x [\pi^2/2 + (\nabla\phi)^2/2 + V(\phi)]$, the short-time matrix element is:

$$\langle \phi_{k+1} | e^{-iH\varepsilon} | \phi_k \rangle = \int \mathcal{D}\pi_k \exp\left(i\varepsilon \int d^3x [\pi_k \dot{\phi}_k - H(\phi_k, \pi_k)]\right) + O(\varepsilon^2), \quad (36.28)$$

where $\dot{\phi}_k = (\phi_{k+1} - \phi_k)/\varepsilon$ is the discrete time derivative. The integral over π_k is a *Gaussian integral*: the integrand is $e^{i\varepsilon \int (\pi \dot{\phi} - \pi^2/2 - \dots)}$, which is Gaussian in π . Completing the square in π :

$$\int \mathcal{D}\pi \exp\left(i\varepsilon \int d^3x \left[\pi \dot{\phi} - \frac{\pi^2}{2}\right]\right) = \mathcal{N} \cdot \exp\left(i\varepsilon \int d^3x \frac{\dot{\phi}^2}{2}\right), \quad (36.29)$$

where $\mathcal{N}$ is a field-independent normalization constant (the determinant of the Gaussian, which is an infinite-dimensional constant and is absorbed into the path integral measure). The result of the Gaussian integral over π is the kinetic term $\dot{\phi}^2/2$, which, combined with the remaining terms from $H(\phi, \pi)$, produces the *Lagrangian*:

$$\mathcal{L}[\phi, \dot{\phi}] = \frac{\dot{\phi}^2}{2} - \frac{(\nabla\phi)^2}{2} - V(\phi) = \frac{1}{2}(\partial_\mu\phi)^2 - V(\phi). \quad (36.30)$$

Repeating this at each time slice and taking $N \rightarrow \infty$ gives the path integral:

$$\langle f | e^{-iHT} | i \rangle = \int \mathcal{D}\phi \exp\left(i \int_0^T d^4x \mathcal{L}[\phi, \partial_\mu\phi]\right). \quad (36.31)$$

The Lagrangian (36.30) has re-emerged from the Hamiltonian by the Legendre transform, performed automatically by the Gaussian integral over the conjugate momentum π . This is the round trip: Lagrangian $\rightarrow$ Hamiltonian (Legendre transform) $\rightarrow$ time slicing $\rightarrow$ Gaussian integral over $\pi \rightarrow$ Lagrangian.

The Gaussian integral and the determinant. The Gaussian integral (36.29) is the path integral version of the finite-dimensional identity:

$$\int d^n\pi \exp\left(-\frac{1}{2}\pi^T M \pi + J^T \pi\right) = \frac{(2\pi)^{n/2}}{\sqrt{\det M}} \exp\left(\frac{1}{2}J^T M^{-1} J\right). \quad (36.32)$$

The determinant $\det M$ of the quadratic form appears in the denominator and contributes a factor $(\det M)^{-1/2}$ to the path integral measure. For the scalar field, M is the identity (the kinetic term $\pi^2/2$ has coefficient matrix $M = 1$), so $\det M = 1$ and the normalization is field-independent — it cancels between numerator and denominator in all physical amplitudes and can be ignored. For more complex Hamiltonians where the kinetic term is $\pi^T M(\phi)\pi/2$ with a field-dependent matrix $M(\phi)$, the determinant $\det M(\phi)$ is field-dependent and *cannot* be ignored: it contributes a non-trivial factor to the path integral that modifies the effective Lagrangian. This is the first glimpse of the Faddeev-Popov determinant: in gauge theories, the restriction of the path integral to the gauge-fixing surface introduces exactly such a field-dependent determinant $\det(\delta G/\delta\alpha)$, which must be handled with care. The Faddeev-Popov procedure exponentiates this determinant using anticommuting ghost fields, converting a field-dependent prefactor into an additional term in the action, exactly as (36.32) converts the determinant of a quadratic form into an exponential.

For fermions: the variables become Grassmann numbers. The round trip for fermions is the same in structure but different in the nature of the integration variables. Fermionic creation and annihilation operators $b^\dagger(\mathbf{k}, s)$ and $b(\mathbf{k}, s)$ satisfy anticommutation relations $\{b(\mathbf{k}, s), b^\dagger(\mathbf{k}', s')\} = \delta^3(\mathbf{k} - \mathbf{k}')\delta_{ss'}$. The coherent state resolution of the identity for fermions requires anticommuting integration variables: *Grassmann numbers* $\psi(\mathbf{x}, t_k)$ satisfying:

$$\psi_\alpha \psi_\beta = -\psi_\beta \psi_\alpha, \quad \psi_\alpha^2 = 0. \quad (36.33)$$

The fermionic path integral:

$$\langle f | e^{-iHT} | i \rangle = \int \mathcal{D}\bar{\psi} \mathcal{D}\psi \exp\left(i \int d^4x \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi\right) \quad (36.34)$$

integrates over Grassmann-valued field configurations $\psi(\mathbf{x}, t)$ and $\bar{\psi}(\mathbf{x}, t)$ at each spacetime point. These are not classical fields: no classical object is Grassmann-valued. They are formal anticommuting variables introduced specifically to handle the fermionic operator algebra in the path integral language.

Why the Lagrangian that re-emerges is not the original one. The Dirac Lagrangian we *started with* was a function of the classical field $\psi_A(x)$ and its spacetime derivatives:

$$\mathcal{L}_{\text{Dirac}}[\psi, \bar{\psi}] = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi, \quad (36.35)$$

where $\psi_A(x)$ is a classical spinor field — an ordinary (commuting) complex number at each spacetime point with a spinor index A . This was the starting point for the Hamiltonian: we Legendre-transformed (36.35) to get H , promoted ψ and $\bar{\psi}$ to operators on Fock space, and derived the Hamiltonian (36.26).

The path integral (36.34) that re-emerges from the time slicing has the *same form* as (36.35) — the same expression $\bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi$ in the exponent — but the variables $\psi(\mathbf{x}, t)$ are now *Grassmann numbers*, not ordinary complex numbers. The form is preserved; the nature of the variables is not. This is a critical distinction that is often glossed over:

- The *original* Lagrangian (36.35) is a function of classical commuting fields: it was a “fictitious classical field” (in Pauli’s language) introduced as a scaffold for constructing the Hamiltonian. Its variables $\psi_A(x)$ are ordinary complex numbers with spinor indices.
- The *path integral* Lagrangian in (36.34) is a function of Grassmann fields: its variables $\psi_A(\mathbf{x}, t)$ are anticommuting formal variables introduced by the fermionic coherent state resolution of the identity. They encode the anticommutation relations of the underlying Fock operators.

The two expressions look identical on the page but describe different mathematical objects. The round trip does not return to the starting point: it returns to a path integral whose integrand has the same functional form as the original Lagrangian but whose variables live in a completely different algebraic structure. This is not a flaw in the formalism; it is an honest reflection of the fact that the fermion field is fundamentally quantum mechanical (it is a Fock operator, not a classical field), and Grassmann numbers are the classical shadow of quantum anticommutativity, introduced to make the path integral well-defined.

The path integral is not a sum over classical field trajectories. The lesson of the round trip is general and applies to both bosons and fermions. The path integral:

$$\langle f | e^{-iHT} | i \rangle = \int \mathcal{D}\phi \exp(iS[\phi]) \quad (36.36)$$

is derived by inserting resolutions of the identity between each time slice of e^{-iHT} . The field $\phi(\mathbf{x}, t)$ at each time slice is the label of the basis state inserted at that slice: it is an integration variable in a matrix multiplication, not a physical field configuration that the quantum system “visits” at time t .

The path $(i, \phi_1, \phi_2, \dots, \phi_{N-1}, f)$ is a sequence of matrix multiplication indices, one per time step, summed over in the matrix product $(e^{-iH\epsilon})^N$. Asking “which path did the field take?” is asking which intermediate index the matrix multiplication passed through — a question that has no physical answer, exactly as asking which intermediate position a particle “was at” between two quantum measurements has no physical answer. The sum over paths is the sum over intermediate matrix indices; the amplitude of each “path” is the product of matrix elements along the index sequence; the full path integral is the matrix power $(e^{-iH\epsilon})^N$ written out in a basis. There are no trajectories, no histories, no classical field configurations that the quantum field explores.

There is only the Fock hidden state h_t , evolving under the Schrödinger equation, and the path integral as a computational technique for evaluating the matrix elements of that evolution.

For bosonic fields, the integration variables $\phi(\mathbf{x}, t_k)$ are ordinary complex numbers — the labels of coherent states — and they superficially resemble classical field values. For fermionic fields, the integration variables $\psi(\mathbf{x}, t_k)$ are Grassmann numbers, making the non-classical nature of the variables explicit. But the conceptual status is the same in both cases: integration labels in a matrix multiplication, not classical field configurations.

Foreshadowing Faddeev-Popov. The Gaussian integral over the conjugate momentum π produced a field-independent determinant in the scalar case. For gauge fields, the analogous step — restricting the path integral to the gauge-fixing surface $G[A] = 0$ — produces a *field-dependent* Jacobian determinant $\det(\delta G/\delta \alpha)$ that cannot be ignored or absorbed into the measure. This determinant has exactly the form of a Gaussian integral over anticommuting variables: by (36.32) applied to anticommuting (Grassmann) fields, a fermionic Gaussian integral produces a determinant in the *numerator* rather than the denominator:

$$\int \mathcal{D}c \mathcal{D}\bar{c} \exp(-\bar{c} M c) = \det M, \quad (36.37)$$

in contrast to the bosonic case $\int \mathcal{D}\phi \exp(-\phi^T M \phi/2) = (\det M)^{-1/2}$. The Faddeev-Popov procedure exploits (36.37): by introducing anticommuting ghost fields c^a and $\bar{c}^a$ whose Gaussian integral reproduces $\det(\delta G/\delta \alpha)$, the field-dependent Jacobian is absorbed into the action as a ghost term, flattening the curved gauge-fixing surface into a flat extended space. The ghost fields are not physical particles; they are the anticommuting integration variables that handle the Jacobian, exactly as the Grassmann variables $\psi(\mathbf{x}, t)$ in the fermionic path integral handle the anticommutativity of the fermionic operators. Both are Grassmann shadows of quantum algebra, introduced to make a path integral computation tractable.

Step	Bosonic field	Fermionic field
Starting point	Lagrangian $\mathcal{L}[\phi]$, commuting ϕ	Lagrangian $\mathcal{L}[\psi, \bar{\psi}]$, commuting ψ
Hamiltonian	$H[\hat{\phi}, \hat{\pi}]$: operator on Fock space	$H[\hat{\psi}, \hat{\bar{\psi}}]$: operator on Fock space
Time slicing	Coherent state basis: commuting $\phi(\mathbf{x}, t_k)$	Coherent state basis: Grassmann $\psi(\mathbf{x}, t_k)$
Conjugate integral	Gaussian over π : gives $(\det M)^{-1/2}$	Gaussian over $\bar{\psi}$: gives $\det M$
Re-emergent Lagrangian	Same form $\mathcal{L}[\phi]$, commuting variables	Same form $\mathcal{L}[\psi]$, Grassmann variables
Path integral variables	Integration labels (not classical trajectories)	Integration labels (not classical trajectories)
Faddeev-Popov ghost	N/A	Same Grassmann integral: $\int \mathcal{D}c \mathcal{D}\bar{c} e^{-\bar{c} M c} = \det M$

36.6 The QED path integral

Quantum electrodynamics describes the interaction of the Dirac electron with the photon field A_μ . The QED action is:

$$S_{\text{QED}}[\bar{\psi}, \psi, A] = \int d^4x \left[\bar{\psi}^A (i\gamma^\mu D_\mu - m)^A_B \psi^B - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \right], \quad (36.38)$$

where $D_\mu = \partial_\mu - ieA_\mu$ is the covariant derivative and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$.

Two kinds of field, two kinds of integral

The QED path integral must integrate over both the photon field A_μ and the electron-positron field ψ . These are qualitatively different objects:

Field	Statistics	Variable type	Measure
$A_\mu(x)$	Bosonic	Ordinary real/complex number	$\mathcal{D}A_\mu$
$\psi^A(x), \bar{\psi}^A(x)$	Fermionic	Grassmann number	$\mathcal{D}\bar{\psi} \mathcal{D}\psi$

The QED path integral is:

$$Z_{\text{QED}} = \int \mathcal{D}A_\mu \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{iS_{\text{QED}}[\bar{\psi}, \psi, A]}. \quad (36.39)$$

This is simultaneously:

- An ordinary functional integral over $A_\mu(x) \in \mathbb{R}$ at each point x and each μ , coming from the bosonic Fock space of the photon.
- A Grassmann functional integral over $\psi^A(x)$ and $\bar{\psi}^A(x)$ at each point x and each Dirac index A , coming from the fermionic Fock space.

Gauge redundancy and the Faddeev–Popov procedure

The path integral (36.39) has an immediate problem: the action S_{QED} is gauge invariant, meaning it is constant on gauge orbits $[A_\mu]$. Integrating over all A_μ therefore counts each physical configuration infinitely many times (once for each gauge transformation). The path integral is formally infinite.

The Faddeev–Popov procedure isolates the integral over physical configurations by factoring out the (infinite) volume of the gauge group. For $U(1)$ QED, one imposes a gauge-fixing condition such as the Lorenz gauge $\partial^\mu A_\mu = 0$ and introduces a gauge-fixing term:

$$S_{\text{QED}} \mapsto S_{\text{QED}} - \frac{1}{2\xi} (\partial^\mu A_\mu)^2, \quad (36.40)$$

where ξ is the gauge parameter ($\xi = 1$: Feynman gauge; $\xi \rightarrow 0$: Lorenz gauge). The gauge-fixed path integral is:

$$Z_{\text{QED}} = \int \mathcal{D}A_\mu \mathcal{D}\bar{\psi} \mathcal{D}\psi \exp \left[i \int d^4x \left(\bar{\psi} (i\not{D} - m) \psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2\xi} (\partial \cdot A)^2 \right) \right]. \quad (36.41)$$

This is now a well-defined functional integral. The gauge-fixing term breaks the manifest gauge invariance of the integrand but preserves all gauge-invariant physical predictions, as guaranteed by the Faddeev–Popov formalism.

Integrating out the fermions. Since the action (36.41) is quadratic in the Grassmann fields $\psi, \bar{\psi}$ (the interaction $e\bar{\psi}A\psi$ is linear in A_μ but bilinear in ψ), the Grassmann integral can be performed exactly:

$$\int \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{i \int \bar{\psi} (i\not{D} - m) \psi} = \det(i\not{D} - m), \quad (36.42)$$

where now $D_\mu = \partial_\mu - ieA_\mu$ contains the background photon field. The QED path integral becomes:

$$Z_{\text{QED}} = \int \mathcal{D}A_\mu \det(i\not{D} - m) e^{i \int (-\frac{1}{4} F^2 - \frac{1}{2\xi} (\partial A)^2)}. \quad (36.43)$$

The fermion determinant $\det(i\not{D} - m)$ is a function of the photon field A_μ . It encodes all quantum effects of the electron in the background of the photon field: expanding it perturbatively in e

generates closed electron loop diagrams. In the quenched approximation (ignoring electron loops), one sets $\det(i\mathcal{D} - m) = 1$; in full QED, the determinant is kept and computed perturbatively.

The Fock-space interpretation of the path integral. In the time-slicing picture, the QED path integral is a product of N matrix elements at each time step, each of the form:

$$\langle A_{k+1}; \eta_{k+1} | e^{-iH_{\text{QED}}\varepsilon} | A_k; \eta_k \rangle, \quad (36.44)$$

where $|A_k; \eta_k\rangle$ is a tensor product of a bosonic coherent state for the photon (eigenvalue $A_k^\mu(\mathbf{x}) \in \mathbb{R}$) and a fermionic coherent state for the electron (eigenvalue $\eta_k^A(\mathbf{x}) \in \text{Grassmann}$). The product over all time steps gives the full evolution amplitude. The path integral is the continuum limit of this N -fold matrix product, with the bosonic measure $\mathcal{D}A_\mu$ arising from the photon coherent states and the Grassmann measure $\mathcal{D}\bar{\psi}\mathcal{D}\psi$ arising from the electron coherent states.

36.7 The electroweak path integral

The electroweak sector of the Standard Model adds the $SU(2)_L$ gauge fields W_μ^a , the $U(1)_Y$ gauge field B_μ , and the Higgs doublet ϕ^i to the QED structure. The action is:

$$S_{\text{EW}} = \int d^4x \left[|D_\mu\phi|^2 - V(\phi) - \frac{1}{4}W_{\mu\nu}^a W^{a\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} + \bar{L}_L(i\mathcal{D})L_{L,i} + \bar{e}_R(i\mathcal{D})e_R - (y_e\bar{L}_{L,i}H^ie_R + \text{h.c.}) \right], \quad (36.45)$$

where D_μ now packages both the $SU(2)_L$ and $U(1)_Y$ covariant derivatives as in (31.34).

The electroweak path integral integrates over all fields:

$$Z_{\text{EW}} = \int \underbrace{\mathcal{D}W_\mu^a \mathcal{D}B_\mu}_{\text{bosonic (ordinary)}} \underbrace{\mathcal{D}\phi^i \mathcal{D}\phi^{i*}}_{\text{bosonic (ordinary)}} \underbrace{\mathcal{D}\bar{L}_L \mathcal{D}L_L \mathcal{D}\bar{e}_R \mathcal{D}e_R}_{\text{fermionic (Grassmann)}} e^{iS_{\text{EW}}}. \quad (36.46)$$

The Higgs field: ordinary complex integration

The Higgs doublet ϕ^i is a spin-0 bosonic field. Its path integral variables $\phi^i(x) \in \mathbb{C}$ are ordinary complex numbers at each spacetime point, exactly as for the scalar field. They are eigenvalues of the bosonic field operator $\hat{\phi}^i(\mathbf{x})$ on bosonic coherent states, and the measure $\mathcal{D}\phi^i \mathcal{D}\phi^{i*}$ is an ordinary functional integral.

This is a crucial point: *the Higgs field is integrated over with an ordinary (non-Grassmann) measure because the Higgs is a boson.* The distinction is entirely determined by the statistics of the field, not by its gauge representation or the complexity of its potential.

Spontaneous symmetry breaking in the path integral. In the path integral (36.46), spontaneous symmetry breaking is not imposed by hand — it emerges from the integral. The potential $V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4$ has a saddle point at $\phi = 0$ (a local maximum for $\mu^2 > 0$) and a circle of minima at $|\phi| = v$. The path integral is dominated by field configurations near these minima: in the semiclassical (small $\hbar$) approximation, the dominant contributions come from $\phi \approx v$.

To make this explicit, shift $\phi^i = \langle\phi^i\rangle + H^i$ where $\langle\phi^i\rangle = v/\sqrt{2}(0,1)^T$ in unitary gauge and H^i is the fluctuation (the physical Higgs field). The action becomes the action of the broken-phase theory, with the Higgs mass term $+m_h^2|H|^2/2$ (positive curvature at the true minimum), the W and Z mass terms from $|D_\mu\phi|^2|_{\phi=v}$, and the fermion mass terms from the Yukawa coupling at $\phi = v$.

Non-Abelian gauge fixing: Faddeev–Popov ghosts

For the $SU(2)_L$ gauge fields W_μ^a , the gauge-fixing procedure is more involved than for $U(1)$. Imposing $\partial^\mu W_\mu^a = 0$ (Lorenz gauge for non-Abelian fields) introduces not just a gauge-fixing term but also *Faddeev–Popov ghost fields* $c^a, \bar{c}^a$:

$$S_{\text{ghost}} = \int d^4x \bar{c}^a (-\partial^\mu D_\mu^{\text{adj}})^{ab} c^b, \quad (36.47)$$

where D_μ^{adj} is the covariant derivative in the adjoint representation and $a, b \in \{1, 2, 3\}$ are $SU(2)$ adjoint indices. The ghost fields $c^a, \bar{c}^a$ are *anticommuting scalar fields*: they are Grassmann numbers but carry no spinor index and are not physical particles. They are a computational artefact of the gauge-fixing procedure, introduced to cancel the unphysical polarisation modes of the W boson in loop diagrams.

Why ghosts must be Grassmann. The Faddeev–Popov determinant from gauge fixing is:

$$\Delta_{\text{FP}}[A] = \det(-\partial^\mu D_\mu^{\text{adj}}). \quad (36.48)$$

This is a determinant (not its inverse), so it must be represented by Grassmann variables (which give $\det M$) rather than ordinary variables (which give $1/\det M$). If we tried to use ordinary (bosonic) scalar fields for the ghost, we would get $1/\det M$ and cancel the wrong things. The Grassmann nature of the ghost fields is not a choice: it is forced by the requirement that the Faddeev–Popov procedure correctly removes the gauge overcounting.

The ghost contribution cancels the unphysical longitudinal and timelike polarisation modes of the W bosons in loop diagrams, ensuring that only the two physical transverse modes (or three for massive $W^\pm, Z^0$ after symmetry breaking) propagate.

36.8 The full Standard Model path integral

Combining all sectors, the full Standard Model path integral is:

$$\begin{aligned} Z_{\text{SM}} = & \int \underbrace{\mathcal{D}G_\mu^a \mathcal{D}W_\mu^i \mathcal{D}B_\mu}_{\text{gauge bosons (ordinary)}} \underbrace{\mathcal{D}\phi^i \mathcal{D}\phi^{i*}}_{\text{Higgs (ordinary)}} \\ & \times \underbrace{\mathcal{D}\bar{c}_G^a \mathcal{D}c_G^a \mathcal{D}\bar{c}_W^i \mathcal{D}c_W^i}_{\text{ghosts (Grassmann scalars)}} \\ & \times \underbrace{\prod_f \mathcal{D}\bar{\Psi}_f \mathcal{D}\Psi_f}_{\text{fermions (Grassmann spinors)}} e^{iS_{\text{SM}}}, \end{aligned} \quad (36.49)$$

where f runs over all fermion species (quarks and leptons, all generations, all colours and isospin components), $a \in \{1, \dots, 8\}$ labels the $SU(3)_c$ ghosts, and $i \in \{1, 2, 3\}$ labels the $SU(2)_L$ ghosts.

The Standard Model action is:

$$\begin{aligned} S_{\text{SM}} = & \int d^4x \left[\underbrace{-\frac{1}{4}G_{\mu\nu}^a G^{a\mu\nu} - \frac{1}{4}W_{\mu\nu}^i W^{i\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}}_{\text{gauge kinetic terms}} \right. \\ & + \underbrace{|D_\mu\phi|^2 - V(\phi)}_{\text{Higgs}} + \underbrace{\sum_f \bar{\Psi}_f i \not{D}_f \Psi_f}_{\text{fermion kinetic + gauge coupling}} \\ & + \underbrace{\sum_{f,f'} (y_{ff'} \bar{\Psi}_{f,L} \phi \Psi_{f',R} + \text{h.c.})}_{\text{Yukawa couplings}} + \underbrace{S_{\text{gf}} + S_{\text{ghost}}}_{\text{gauge fixing + ghosts}} \left. \right], \end{aligned} \quad (36.50)$$

where $\mathcal{D}_f$ is the covariant derivative appropriate to the gauge representation of species f (including colour, isospin, and hypercharge couplings as in (30.28)).

The complete field content and variable types

Field	Spin	Stat.	Variable type	Indices
$G_\mu^a(x)$	1	Boson	Ordinary real	$\mu, a \in \{1, \dots, 8\}$
$W_\mu^i(x)$	1	Boson	Ordinary real	$\mu, i \in \{1, 2, 3\}$
$B_\mu(x)$	1	Boson	Ordinary real	μ
$\phi^i(x)$	0	Boson	Ordinary complex	$i \in \{1, 2\}$
$c_G^a(x), \bar{c}_G^a(x)$	0	Ghost	Grassmann scalar	$a \in \{1, \dots, 8\}$
$c_W^i(x), \bar{c}_W^i(x)$	0	Ghost	Grassmann scalar	$i \in \{1, 2, 3\}$
$Q_L^{i\alpha}(x)$	$\frac{1}{2}$	Fermion	Grassmann spinor	A, i, α
$u_R^\alpha(x)$	$\frac{1}{2}$	Fermion	Grassmann spinor	A, α
$d_R^\alpha(x)$	$\frac{1}{2}$	Fermion	Grassmann spinor	A, α
$L_L^i(x)$	$\frac{1}{2}$	Fermion	Grassmann spinor	A, i
$e_R(x)$	$\frac{1}{2}$	Fermion	Grassmann spinor	A

Every row in this table is an integration variable in (36.49). Each variable is the *eigenvalue of the corresponding field operator in a Fock-space coherent state*, introduced at each time slice to resolve the identity. The distinction between ordinary and Grassmann variables tracks the boson/fermion distinction exactly.

The path integral as a matrix product over time slices

The Standard Model path integral (36.49) is derived by slicing the time evolution operator $e^{-iH_{SM}T}$ into N steps. At each step k , one inserts the full Standard Model resolution of the identity:

$$\begin{aligned}
1_{SM} = & \int \left(\prod_\mu \mathcal{D}G_{\mu,k}^a \mathcal{D}W_{\mu,k}^i \mathcal{D}B_{\mu,k} \right) \left(\mathcal{D}\phi_k^i \mathcal{D}\phi_k^{i*} \right) \\
& \times \left(\prod_f d\bar{\eta}_{f,k} d\eta_{f,k} e^{-\bar{\eta}_{f,k} \eta_{f,k}} \right) \\
& \times |G_k; W_k; B_k; \phi_k; \eta_k\rangle \langle G_k; W_k; B_k; \phi_k; \eta_k|,
\end{aligned} \tag{36.51}$$

where:

- $|G_k; W_k; B_k\rangle$ are bosonic coherent states for the gauge fields with ordinary real eigenvalues $G_{\mu,k}^a(\mathbf{x}), W_{\mu,k}^i(\mathbf{x}), B_{\mu,k}(\mathbf{x})$.
- $|\phi_k\rangle$ is a bosonic coherent state for the Higgs with ordinary complex eigenvalue $\phi_k^i(\mathbf{x}) \in \mathbb{C}$.
- $|\eta_{f,k}\rangle$ are fermionic coherent states for each fermion species f with Grassmann eigenvalues $\eta_{f,k}^A(\mathbf{x})$.

The N -fold matrix product of matrix elements:

$$\prod_{k=0}^{N-1} \langle \text{all fields}_{k+1} | e^{-iH_{SM}\varepsilon} | \text{all fields}_k \rangle \tag{36.52}$$

gives, in the limit $N \rightarrow \infty$, the Standard Model path integral (36.49) with $e^{iS_{\text{SM}}}$ as the weight.

Each factor in (36.52) is a transition amplitude between adjacent Fock-space states at consecutive time slices. The path integral is the sum (integral) over all possible intermediate field configurations at each time slice, weighted by the phase $e^{iS_{\text{SM}}}$. This is the precise statement of the path integral as a continuous matrix product: the amplitude for the universe to evolve from one Fock-space configuration to another is the sum over all possible intermediate Fock-space configurations, with each history weighted by $e^{iS_{\text{SM}}}$.

Perturbative expansion and Feynman rules

Splitting the action as $S_{\text{SM}} = S_0 + S_{\text{int}}$ (free plus interaction), the path integral becomes:

$$Z_{\text{SM}} = \int \mathcal{D}[\text{all fields}] e^{iS_0} e^{iS_{\text{int}}} = Z_0 \langle e^{iS_{\text{int}}} \rangle_0, \quad (36.53)$$

where $\langle \cdots \rangle_0$ denotes the expectation value in the free theory. Expanding $e^{iS_{\text{int}}}$ as a power series in the coupling constants (e, g, g', g_s, y_f):

$$e^{iS_{\text{int}}} = 1 + iS_{\text{int}} + \frac{1}{2!}(iS_{\text{int}})^2 + \cdots \quad (36.54)$$

Each term in this expansion involves a product of interaction vertices from (36.50). Evaluating the Gaussian integrals over the free fields, each pairing of fields $\bar{\psi}^A(x)\psi^B(y)$, $A_\mu(x)A_\nu(y)$, etc. under the free measure gives a propagator:

$$\langle \psi^A(x)\bar{\psi}^B(y) \rangle_0 = \frac{i(\not{\partial} + m)^{AB}}{-\partial^2 - m^2 + i\epsilon} \xrightarrow{\text{Fourier}} \frac{i(\not{p} + m)^{AB}}{p^2 - m^2 + i\epsilon} \quad (\text{fermion propagator}), \quad (36.55)$$

$$\langle A_\mu(x)A_\nu(y) \rangle_0 = \frac{-i\eta_{\mu\nu}}{-\partial^2 + i\epsilon} \xrightarrow{\text{Fourier}} \frac{-i\eta_{\mu\nu}}{k^2 + i\epsilon} \quad (\text{photon propagator, Feynman gauge}). \quad (36.56)$$

These are the *Feynman propagators*: the Green's functions of the free field equations. For the Grassmann (fermion) fields, the pairing $\langle \psi\bar{\psi} \rangle_0$ comes from the Grassmann Gaussian integral; the sign rules for fermion loops (factors of -1 for each closed fermion loop in a Feynman diagram) are direct consequences of the Grassmann algebra.

Fermion loops and the minus sign. A closed fermion loop in a Feynman diagram involves a trace over Dirac indices and a cyclic permutation of Grassmann fields around the loop. Each cyclic permutation of two Grassmann variables introduces a minus sign (from the anticommutativity $\psi\bar{\psi} = -\bar{\psi}\psi$). The overall sign for a loop with n fermion propagators involves n such permutations; the net sign is $(-1)^n(-1) = (-1)^{n+1}$. The conventional statement is simply: each closed fermion loop contributes a factor of -1 . This is a pure Grassmann algebra effect with no bosonic analogue.

Summary: what the Standard Model path integral is.

Path integral concept	Precise meaning
Integration variable $\phi(x)$ for boson	Eigenvalue of $\hat{\phi}(\mathbf{x})$ on bosonic coherent state $ \phi\rangle$; ordinary complex number
Integration variable $\psi(x)$ for fermion	Eigenvalue of $\hat{\Psi}(\mathbf{x})$ on fermionic coherent state $ \eta\rangle$; Grassmann number
Measure $\mathcal{D}\phi$	Product of ordinary integrals $\prod_x d\phi(x)$; from bosonic resolution of identity
Measure $\mathcal{D}\bar{\psi}\mathcal{D}\psi$	Product of Berezin integrals $\prod_x d\bar{\psi}(x)d\psi(x)$; from fermionic resolution of identity
e^{iS} weight	Product of N matrix elements $\langle k+1 e^{-iH\varepsilon} k\rangle$ over time slices
$\int \mathcal{D}[\text{all}] e^{iS}$	Continuum limit of N -fold matrix product in Fock space; sums over all intermediate Fock configurations
Propagator $\langle\phi(x)\phi(y)\rangle_0$	Free-field two-point function; from Gaussian integral over the free action
Vertex from $\mathcal{L}_{\text{int}}$	Off-diagonal block of H_{int} in Fock basis; transfers amplitude between Fock configurations
Fermion loop $\rightarrow$ factor -1	Grassmann anticommutativity under cyclic permutation around the loop
Ghost fields $c^a, \bar{c}^a$	Grassmann scalars implementing Faddeev–Popov determinant; cancel unphysical gauge boson polarisations

The Standard Model path integral (36.49) is not an integral over classical field configurations. It is the continuum limit of a product of quantum mechanical transition amplitudes between consecutive Fock-space states, with the bosonic states labelled by ordinary numbers and the fermionic states labelled by Grassmann numbers. The weight $e^{iS_{\text{SM}}}$ is the accumulated phase from all these transitions. The Feynman rules — propagators and vertices — are the perturbative expansion of this matrix product in powers of the coupling constants. The path integral is the source code; the Feynman rules are the compiled output.

37 Gauge Orbits, Faddeev–Popov, and BRST

37.1 Overview: dimension reduction and augmentation

The path integral over gauge fields overcounts: it integrates over every gauge field configuration, including all configurations related by local gauge transformations, which are physically identical. The physical configuration space is not the full field space $\mathcal{A}$ but the quotient $\mathcal{A}/\mathcal{G}$ — the space of gauge orbits. The challenge is to restrict the path integral to this physical space without breaking the symmetries that make the theory tractable. The Faddeev–Popov and BRST procedures solve this challenge in three sequential steps, each with a precise linear-algebraic interpretation.

Step 1 — Gauge fixing: nonlinear dimension reduction. A gauge-fixing condition $G[A] = 0$ (for example the Lorenz condition $\partial^\mu A_\mu = 0$) selects one representative field configuration from each gauge orbit. This is a *nonlinear dimension reduction*: the condition $G[A] = 0$ defines a curved surface in the infinite-dimensional field space $\mathcal{A}$, and restricting the path integral to this surface cuts the integration down from $\mathcal{A}$ to the physical configuration space $\mathcal{A}/\mathcal{G}$. The problem is that this surface is curved, not flat: changing variables from $\mathcal{A}$ to the gauge-fixing surface introduces a Jacobian $\det(\delta G/\delta \alpha)$ that depends on the field A_μ and cannot be ignored. Dropping it gives the wrong answer; keeping it as a prefactor makes the path integral intractable.

Step 2 — Faddeev–Popov: dimension augmentation to flatten the space. The Faddeev–Popov trick absorbs the Jacobian into the action by introducing anticommuting ghost fields c^a and $\bar{c}^a$:

$$\det\left(\frac{\delta G}{\delta \alpha}\right) = \int \mathcal{D}c \mathcal{D}\bar{c} \exp\left(-\int d^4x \bar{c}^a \frac{\delta G^a}{\delta \alpha^b} c^b\right). \quad (37.1)$$

This is a *dimension augmentation*: new field variables (the ghosts) are added to the integration, specifically chosen so that their Grassmann integral reproduces the Jacobian exactly. The effect is to *flatten the field space*: the combined space of gauge fields plus ghosts $\mathcal{A} \times \mathcal{G}_{\text{ghost}}$ is treated as a flat linear space, with the curvature of the gauge-fixing surface absorbed into the ghost action rather than appearing as a prefactor. The nonlinear dimension reduction of Step 1 has been traded for a flat-space integration over an extended field space, at the cost of introducing ghost fields. The two operations — dimension reduction (removing gauge directions) and dimension augmentation (adding ghost directions) — cancel each other’s formal effect on the path integral, leaving the correct physical answer.

Step 3 — BRST: linear dimension reduction to the physical subspace. The gauge-fixed action (Yang–Mills plus gauge-fixing plus ghosts) has a residual nilpotent symmetry: the BRST operator $\mathcal{Q}$, satisfying $\mathcal{Q}^2 = 0$. The physical Hilbert space is the BRST cohomology:

$$\mathcal{H}_{\text{phys}} = \ker \mathcal{Q} / \text{im } \mathcal{Q}. \quad (37.2)$$

This is a *linear dimension reduction*: $\ker \mathcal{Q}$ is the linear subspace of states annihilated by $\mathcal{Q}$ (the null space of $\mathcal{Q}$), and $\text{im } \mathcal{Q}$ is the linear subspace of states of the form $\mathcal{Q}|\chi\rangle$ (the column space of $\mathcal{Q}$). The quotient identifies all states that differ by a BRST-exact term, retaining only the states that are in the null space but not in the column space. The nilpotency $\mathcal{Q}^2 = 0$ guarantees $\text{im } \mathcal{Q} \subset \ker \mathcal{Q}$, so the quotient is well-defined. The unphysical degrees of freedom — ghost fields, longitudinal gauge modes, temporal polarisations — are either outside $\ker \mathcal{Q}$ or inside $\text{im } \mathcal{Q}$; only the transverse, gauge-invariant physical states survive in $\mathcal{H}_{\text{phys}}$.

The complete picture. The three steps together convert a fundamentally nonlinear geometric problem — integrating over the space of gauge orbits $\mathcal{A}/\mathcal{G}$ — into a sequence of linear algebraic operations:

Step	Operation	Type	Price
Gauge fixing	Select one representative per orbit: $G[A] = 0$	Nonlinear dimension reduction	Jacobian $\det(\delta G/\delta \alpha)$
Faddeev-Popov	Exponentiate Jacobian as ghost action	Dimension augmentation; flatten the space	Ghost fields $c, \bar{c}$ added
BRST cohomology	Quotient $\ker \mathcal{Q}/\text{im } \mathcal{Q}$	Linear dimension reduction; null space mod column space	None: exact linear subspace quotient

The gauge orbit structure is a nonlinear geometry; the three steps linearise it. Faddeev-Popov flattens the curved gauge-fixing surface into a flat extended space by absorbing the curvature into the ghost action. BRST then performs the final linear dimension reduction, projecting onto the physical subspace by the null-space-mod-column-space construction that is familiar from linear algebra and homological algebra alike. The result is a well-defined, gauge-invariant, positive-norm physical Hilbert space — exactly the Fock space of transverse gauge bosons and physical matter fields that the theory predicts.

37.2 The overcounting problem

The naive Standard Model path integral:

$$Z_{\text{SM}} = \int \mathcal{D}\Psi \mathcal{D}G \mathcal{D}W \mathcal{D}B \mathcal{D}\phi e^{iS_{\text{SM}}} \quad (37.3)$$

is gauge-invariant: S_{SM} is constant on gauge orbits. Every field configuration A_μ and its gauge transform $A_\mu^U = U A_\mu U^\dagger + \frac{i}{g}(\partial_\mu U)U^\dagger$ contribute the same weight $e^{iS_{\text{SM}}}$. Integrating over all A_μ therefore counts each physical state once for every gauge transformation $U \in \mathcal{G}$, so:

$$Z_{\text{naive}} = \text{vol}(\mathcal{G}) \times Z_{\text{phys}}, \quad (37.4)$$

where $\text{vol}(\mathcal{G})$ is the volume of the infinite-dimensional gauge group — infinite, and meaningless as a prefactor.

The gauge orbit. For any configuration A_μ , its *gauge orbit* is:

$$[A_\mu] = \{A_\mu^U : U(x) \in \mathcal{G}\}. \quad (37.5)$$

All configurations on the same orbit are physically identical. The physical configuration space is the quotient $\mathcal{A}/\mathcal{G}$: the space of orbits.

The onion picture. Think of $\mathcal{A}$ as an onion. Each layer is one gauge orbit: configurations that are gauge-equivalent to each other. Moving *within* a layer is a gauge transformation (unphysical); moving *between* layers corresponds to genuinely different physical states. Physical observables depend only on which layer you are on, not where within the layer.

The naive path integral integrates over every point of every layer. What we want is to integrate *once per layer*. To do this, we impose a gauge condition $G^a[A] = 0$ that selects one representative from each orbit — a slice through the onion — and account for the Jacobian of this selection.

The (a_0, a_1) decomposition. Every field configuration can be decomposed as:

$$A_\mu \leftrightarrow (a_0[A], a_1[A]), \quad (37.6)$$

where $a_0 \in \mathcal{G}$ parametrises the position *within* the gauge orbit and $a_1 \in \mathcal{A}/\mathcal{G}$ labels the orbit itself. Gauge transformations change a_0 while fixing a_1 ; physical observables depend on a_1 alone.

Explicit example: QED. In QED, the decomposition is the Helmholtz decomposition of the 3-vector potential:

$$A_i(\mathbf{x}) = A_i^\perp(\mathbf{x}) + \partial_i \chi(\mathbf{x}), \quad (37.7)$$

where $\nabla \cdot \mathbf{A}^\perp = 0$ (transverse, physical) and $\partial_i \chi$ is pure gauge. Under $A_i \rightarrow A_i + \partial_i \alpha$, the transverse part $A_i^\perp$ is unchanged and $\chi \rightarrow \chi + \alpha$. So $a_1 = A^\perp$ (the physical part, two polarisations for each $\mathbf{k}$) and $a_0 = \chi$ (the gauge part, one function of $\mathbf{x}$). The naive measure factorises:

$$\mathcal{D}A_\mu = \mathcal{D}A^\perp \mathcal{D}A^L \mathcal{D}A^0 = \underbrace{\mathcal{D}\chi}_{\text{vol}(U(1))=\infty} \times \underbrace{\mathcal{D}A^\perp}_{\text{physical}}. \quad (37.8)$$

The infinite $\mathcal{D}\chi$ factor is the overcounting. For non-Abelian theories, the orbit is curved and the decomposition has no closed-form expression — which is why we need the Faddeev–Popov procedure rather than a simple projection.

37.3 The Faddeev–Popov procedure

Step 1: insert a 1 written as a gauge-fixed delta function

The key identity, due to Faddeev and Popov, inserts a factor of 1 into the path integral in a cleverly chosen form. For any smooth gauge-fixing function $G^a[A]$ (e.g. $G^a[A] = \partial^\mu A_\mu^a$ for Lorenz gauge), define the Faddeev–Popov determinant $\Delta_{FP}[A]$ by:

$$1 = \Delta_{FP}[A] \int \mathcal{D}U \delta^{(4)}(G^a[A^U]), \quad (37.9)$$

where $\mathcal{D}U$ is the (Haar) measure on the gauge group and A^U is the gauge transform of A by $U(x)$. This says: integrating the delta function $\delta(G[A^U])$ over all gauge transformations U , weighted by $\Delta_{FP}[A]$, gives exactly 1.

The Faddeev–Popov determinant $\Delta_{FP}[A]$ is the Jacobian of the map $U \mapsto G[A^U]$ at the point where $G[A^U] = 0$. For the Lorenz gauge $G^a = \partial^\mu A_\mu^a$:

$$\Delta_{FP}[A] = \det M[A], \quad M^{ab}(x, y) = \left. \frac{\delta G^a[A^U](x)}{\delta \alpha^b(y)} \right|_{G=0} = -\partial^\mu (D_\mu)^{ab} \delta^4(x - y), \quad (37.10)$$

where $(D_\mu)^{ab} = \partial_\mu \delta^{ab} + g f^{acb} A_\mu^c$ is the covariant derivative in the adjoint representation. For QED ($f^{abc} = 0$): $M^{ab} = -\partial^2 \delta^{ab}$, which is field-independent and can be pulled out of the path integral as an overall constant. This is why QED ghosts decouple.

Step 2: insert the identity into the path integral

Inserting (37.9) into (37.3):

$$\begin{aligned} Z_{\text{naive}} &= \int \mathcal{D}A e^{iS[A]} \Delta_{FP}[A] \int \mathcal{D}U \delta(G[A^U]) \\ &= \int \mathcal{D}U \int \mathcal{D}A e^{iS[A]} \Delta_{FP}[A] \delta(G[A^U]). \end{aligned} \quad (37.11)$$

Now change variables $A \rightarrow A^{U^{-1}}$ in the inner integral (a gauge transformation, which leaves $\mathcal{D}A$ and $S[A]$ invariant, and transforms $\Delta_{FP}[A] \rightarrow \Delta_{FP}[A^{U^{-1}}] = \Delta_{FP}[A]$ since the FP determinant is gauge-invariant):

$$Z_{\text{naive}} = \underbrace{\int \mathcal{D}U}_{\text{vol}(\mathcal{G})} \underbrace{\int \mathcal{D}A e^{iS[A]} \Delta_{FP}[A] \delta(G[A])}_{Z_{\text{phys}}}. \quad (37.12)$$

The gauge group volume $\text{vol}(\mathcal{G})$ factorises and can be divided out. What remains is the physical path integral:

$$Z_{\text{phys}} = \int \mathcal{D}A \, e^{iS[A]} \Delta_{FP}[A] \delta(G[A]). \quad (37.13)$$

This integrates over one representative per gauge orbit (the δ enforces the gauge condition) with the correct Jacobian weight $\Delta_{FP}[A]$.

Step 3: soften the delta function to a Gaussian

The sharp delta function $\delta(G[A])$ is inconvenient for Feynman rules. Soften it by averaging over gauge conditions weighted by a Gaussian in ξ :

$$\delta(G^a[A]) \longrightarrow \exp\left(-\frac{i}{2\xi} \int d^4x (G^a[A])^2\right). \quad (37.14)$$

For Lorenz gauge $G^a = \partial^\mu A_\mu^a$, this adds the gauge-fixing term:

$$\mathcal{L}_{\text{gf}} = -\frac{1}{2\xi} (\partial^\mu A_\mu^a)^2 \quad (37.15)$$

to the Lagrangian. The parameter ξ is the gauge parameter; all physical observables are independent of ξ . Common choices: $\xi = 1$ (Feynman gauge), $\xi = 0$ (Lorenz gauge), $\xi \rightarrow \infty$ (unitary gauge).

Step 4: represent $\det M$ as a Grassmann integral

The Faddeev–Popov determinant $\det M[A]$ in (37.13) is a functional determinant — an infinite-dimensional analogue of $\det M$ for a finite matrix. It depends on the gauge field A_μ through the covariant derivative in (37.10). We cannot simply evaluate it; we need to represent it as a path integral so that Feynman rules can be derived.

From the Grassmann Gaussian formula (36.15):

$$\int \mathcal{D}\bar{c} \mathcal{D}c \, e^{-i \int d^4x \bar{c}^a M^{ab}[A] c^b} = \det M[A]. \quad (37.16)$$

This requires the integration variables c^a , $\bar{c}^a$ to be Grassmann (anticommuting): bosonic variables would give $(\det M)^{-1}$, not $\det M$. The ghost fields c^a and $\bar{c}^a$ are therefore Grassmann-valued scalar fields with adjoint colour index a , with the ghost action:

$$S_{\text{ghost}}[A, c, \bar{c}] = \int d^4x \bar{c}^a (-\partial^\mu D_\mu)^{ab} c^b. \quad (37.17)$$

Expanding the covariant derivative $(D_\mu)^{ab} = \partial_\mu \delta^{ab} + g f^{acb} A_\mu^c$:

$$\mathcal{L}_{\text{ghost}} = \bar{c}^a (-\partial^2) \delta^{ab} c^b + g f^{abc} (\partial^\mu \bar{c}^a) A_\mu^c c^b. \quad (37.18)$$

The first term is the ghost kinetic term (propagator i/k^2); the second is the ghost-gluon interaction vertex (coupling $\sim g f^{abc} p^\mu$).

The full gauge-fixed path integral. Combining all four steps, the physical path integral is:

$$Z_{\text{phys}} = \int \mathcal{D}A \mathcal{D}\bar{c} \mathcal{D}c \, \exp\left(i \int d^4x [\mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{gf}} + \mathcal{L}_{\text{ghost}}]\right). \quad (37.19)$$

This is the well-defined, non-overcounting version of the Standard Model path integral. All overcounting has been removed; the gauge-fixing term $\mathcal{L}_{\text{gf}}$ selects one representative per orbit; the ghost action $\mathcal{L}_{\text{ghost}}$ accounts for the Jacobian of that selection.

ML analogy: the Faddeev–Popov procedure as marginalisation with an auxiliary variable. In variational inference, one faces the same problem: a joint distribution $p(x, z) \propto e^{-E(x)}$ (energy independent of z , so z is redundant) gives an intractable marginal $p(x) = \int dz p(x, z)$ because the integral over z diverges. The solution is to introduce auxiliary variables that render the marginalisation tractable. The Faddeev–Popov procedure is exactly this: z is the gauge orbit coordinate a_0 , the redundant integral is $\int \mathcal{D}U$, the auxiliary variables are the ghosts $c^a, \bar{c}^a$, and the “tractable marginalisation” is the Gaussian ghost path integral that implements $\det M[A]$.

ML concept	Faddeev–Popov concept
Redundant latent variable z	Gauge orbit coordinate a_0
Divergent $\int dz$	$\text{vol}(\mathcal{G}) = \infty$
Auxiliary variable trick	Ghost fields $c^a, \bar{c}^a$
Jacobian $ \partial z / \partial x $	FP determinant $\det M[A]$
Tractable Gaussian marginal	Ghost path integral $\int \mathcal{D}\bar{c} \mathcal{D}c e^{i\bar{c} M c}$
Residual symmetry after marginalisation	BRST symmetry

37.4 Gauge fixing conventions

Different gauge conditions give different Feynman rules but the same physical predictions. The most common choices are:

Gauge	Condition $G^a[A]$	ξ	Lorentz cov.?	Best for
Lorenz	$\partial^\mu A_\mu = 0$	any	Yes	Covariant Feynman rules
Feynman	$\partial^\mu A_\mu = 0$	$\xi = 1$	Yes	Loop calculations
Landau	$\partial^\mu A_\mu = 0$	$\xi \rightarrow 0$	Yes	Ward identities
Coulomb	$\nabla \cdot \mathbf{A} = 0$	—	No	Non-relativistic QED
Axial	$A_3 = 0$	—	No	Lightcone physics
Unitary	Goldstone phase set to 0	$\xi \rightarrow \infty$	Yes	Physical spectrum

For electroweak calculations, the R_ξ gauges are standard. The gauge-fixing condition for the W boson is chosen to cancel the $m_W W^\mu \partial_\mu \pi$ mixing term between the W boson and the would-be Goldstone boson π :

$$G^a[W, \pi] = \partial^\mu W_\mu^a - \xi m_W \pi^a, \quad (37.20)$$

giving the W propagator in R_ξ gauge:

$$\tilde{\Delta}_W^{\mu\nu}(k) = \frac{-i \left(\eta^{\mu\nu} - (1 - \xi) \frac{k^\mu k^\nu}{k^2 - \xi m_W^2} \right)}{k^2 - m_W^2 + i\epsilon}. \quad (37.21)$$

The ξ -dependent pole at $k^2 = \xi m_W^2$ is the Goldstone boson propagator, which cancels against Goldstone and ghost poles in all physical amplitudes, as guaranteed by BRST. In the limit $\xi \rightarrow \infty$ (unitary gauge), the Goldstone poles go to infinity and decouple; only the physical W polarisations remain.

37.5 BRST symmetry

After gauge fixing, the original gauge symmetry is broken — the gauge-fixing term $\mathcal{L}_{\text{gf}}$ is not gauge invariant. However, the gauge-fixed action has a new global fermionic symmetry called *BRST symmetry* (after Becchi, Rouet, Stora, and Tyutin).

Where the BRST transformation comes from

The BRST transformation is a gauge transformation with the ghost field c^a as the gauge parameter, multiplied by a Grassmann number ε :

$$\delta_{\text{BRST}} A_\mu^a = \varepsilon (D_\mu c)^a = \varepsilon (\partial_\mu c^a + g f^{abc} A_\mu^b c^c). \quad (37.22)$$

This is exactly the infinitesimal gauge transformation $A_\mu^a \rightarrow A_\mu^a + D_\mu \alpha^a$ with $\alpha^a \rightarrow \varepsilon c^a$. The reason the ghost appears as the gauge parameter is that the ghost field c^a was introduced to parametrise infinitesimal gauge transformations in the Faddeev–Popov procedure: it is the natural “infinitesimal gauge direction” in the extended field space.

The transformation of the ghost field is forced by the requirement that $\delta_{\text{BRST}}^2 = 0$ (nilpotency):

$$\delta_{\text{BRST}} c^a = -\frac{1}{2} \varepsilon g f^{abc} c^b c^c. \quad (37.23)$$

The product $c^b c^c$ is Grassmann-antisymmetric: $c^b c^c = -c^c c^b$. The factor f^{abc} is totally antisymmetric in a, b, c . So $f^{abc} c^b c^c$ is symmetric under $b \leftrightarrow c$, which combined with the Grassmann antisymmetry gives zero unless the f^{abc} structure supplies the antisymmetry — which it does via the Jacobi identity. This is what ensures $\delta^2 c^a = 0$.

The anti-ghost $\bar{c}^a$ and the Nakanishi–Lautrup auxiliary field B^a transform as:

$$\delta_{\text{BRST}} \bar{c}^a = \varepsilon B^a, \quad \delta_{\text{BRST}} B^a = 0. \quad (37.24)$$

For matter fields (fermions ψ , Higgs ϕ), the BRST transformation is a gauge transformation with parameter εc^a :

$$\delta_{\text{BRST}} \psi = -i\varepsilon g T^a c^a \psi, \quad (37.25)$$

$$\delta_{\text{BRST}} \phi = -i\varepsilon g \left(\frac{\tau^a}{2}\right) c^a \phi. \quad (37.26)$$

Nilpotency: $\delta_{\text{BRST}}^2 = 0$. Applying δ_{BRST} twice:

$$\delta_{\text{BRST}}^2 A_\mu^a = \delta_{\text{BRST}} (\varepsilon (D_\mu c)^a) = \varepsilon^2 \left(D_\mu \left(-\frac{1}{2} g f^{bca} c^b c^c \right) + g f^{abc} (D_\mu c)^b c^c \right) = 0, \quad (37.27)$$

where the last equality follows from the Jacobi identity $f^{abe} f^{ecd} + f^{bce} f^{ead} + f^{cae} f^{ebd} = 0$ and Grassmann anticommutativity $\varepsilon^2 = 0$. Similarly $\delta^2 c^a = 0$, $\delta^2 \psi = 0$, etc.

Nilpotency $\delta^2 = 0$ is the algebraic statement that BRST is a differential: there are no “sub-orbits” within a gauge orbit. It makes BRST behave exactly like the exterior derivative d in differential geometry, with ghost number playing the role of form degree:

Differential geometry	BRST
Exterior derivative d , $d^2 = 0$	BRST operator δ , $\delta^2 = 0$
Form degree p	Ghost number (A_μ^a : 0, c^a : +1, $\bar{c}^a$: -1)
d : degree $p \rightarrow p + 1$	δ : ghost number $n \rightarrow n + 1$
de Rham cohomology $H^p(d)$	BRST cohomology $H^n(\delta)$
Closed forms = exact forms	Physical states: closed but not exact

Remark. [Nilpotency and the boundary of a boundary] The nilpotency $\delta^2 = 0$ of the BRST operator has a beautiful geometric interpretation: *the boundary of a boundary is empty*.

Consider a line segment: its boundary consists of two endpoints. Now consider a disk: its boundary is a circle. But the circle itself has no boundary — it is a closed loop with no endpoints. The boundary of the boundary is empty: $\partial(\partial(\text{disk})) = \partial(\text{circle}) = \emptyset$.

The same is true one dimension higher. The boundary of a ball is a sphere. The sphere is a closed surface with no rim, no edge, no boundary of its own. Again: $\partial(\partial(\text{ball})) = \partial(\text{sphere}) = \emptyset$.

This is a general theorem of topology: $\partial^2 = 0$. The boundary operator ∂ , applied twice, always gives zero. The BRST operator δ is the field-theoretic realisation of exactly this structure. It acts as an exterior derivative along the gauge orbit direction, with ghost number playing the role of form degree. The nilpotency $\delta^2 = 0$ is then the statement that there is no “sub-orbit within a gauge orbit”: once you have moved along the orbit by a gauge transformation (one application of δ), applying δ again finds no further boundary to move along. The gauge orbit is, in this sense, a closed surface — it has no boundary of its own — and the BRST cohomology $\ker \delta / \text{im } \delta$ is the field-theory analogue of the homology groups that classify which cycles are boundaries and which are not.

BRST-exactness of the gauge-fixing sector. The entire gauge-fixing and ghost contribution is BRST-exact:

$$\mathcal{L}_{\text{gf}} + \mathcal{L}_{\text{ghost}} = \delta_{\text{BRST}} \left[\bar{c}^a (G^a[A] + \tfrac{\xi}{2} B^a) \right]. \quad (37.28)$$

Since $\delta^2 = 0$, any BRST-exact term contributes zero to the expectation value of any BRST-closed observable. Changing the gauge choice changes $G^a[A]$, which changes the BRST-exact term in (37.28), but leaves all BRST-closed observables unchanged. *Physical predictions are gauge-independent because they are BRST-cohomology classes, not individual field configurations.*

37.6 Physical states as BRST cohomology

The nilpotent BRST charge Q (the conserved charge corresponding to the BRST symmetry current) satisfies $Q^2 = 0$. This makes Q a differential on the Fock space, and we can define its cohomology.

The physical Hilbert space is:

$$\mathcal{H}_{\text{phys}} = \frac{\ker Q}{\text{im } Q} = \frac{\{\text{states } |\Phi\rangle : Q|\Phi\rangle = 0\}}{\{Q|\chi\rangle : |\chi\rangle \in \mathcal{F}_{\text{ext}}\}}. \quad (37.29)$$

Physical states must be:

- **BRST-closed:** $Q|\Phi\rangle = 0$. They are annihilated by the BRST charge.
- **Not BRST-exact:** $|\Phi\rangle \neq Q|\chi\rangle$ for any $|\chi\rangle$. They are not themselves the BRST transform of something else.
- **Ghost-number zero:** $N_{\text{ghost}}|\Phi\rangle = 0$. Physical states have no net ghost excitations.

The content of BRST cohomology for each species:

State	BRST status	Physical?	Reason
Transverse photon ($\lambda = \pm 1$)	Closed, not exact	Yes	Physical polarisation
Longitudinal photon	Exact ($= Q \text{ghost}\rangle$)	No	Pure gauge
Timelike photon	Not closed	No	Negative norm
Ghost $c^a, \bar{c}^a$ excitation	Ghost number $\neq 0$	No	Not in cohomology
Colour-singlet hadron	Closed, not exact	Yes	Gauge-invariant
Isolated quark	Not closed (Q rotates colour)	No	Carries free colour
Higgs boson h	Closed, not exact	Yes	Physical fluctuation
Goldstone π^i (Lorenz gauge)	Exact	No	Eaten by $W^\pm, Z^0$
Transverse $W^\pm, Z^0$	Closed, not exact	Yes	Physical

Remark. [BRST cohomology as a quotient of subspaces] The BRST cohomology $\ker Q / \text{im } Q$ has a concrete linear-algebraic interpretation that makes the structure transparent.

Think of the BRST operator Q as a sequence of two rectangular matrices Q_1 and Q_2 acting on spaces of increasing ghost number:

$$\cdots \xrightarrow{Q_1} V_1 \xrightarrow{Q_2} V_2 \rightarrow \cdots \quad (37.30)$$

where Q_1 is, say, 20×10 and Q_2 is 10×20 (the sizes reflect the dimensions of the spaces at each ghost number). The nilpotency condition $Q_2 Q_1 = 0$ says: the column space of Q_1 (the image $\text{im } Q_1$, the set of all vectors of the form $Q_1 \chi$) lies entirely within the null space of Q_2 (the kernel $\ker Q_2$, the set of all vectors annihilated by Q_2). In matrix language: every column of Q_1 is orthogonal to every row of Q_2 .

Now consider a vector $\Phi \in V_1$. There are three cases:

- $Q_2 \Phi \neq 0$: Φ is not in $\ker Q_2$. It is not BRST-closed and is unphysical.
- $Q_2 \Phi = 0$ and $\Phi = Q_1 \chi$ for some χ : Φ is in $\ker Q_2$ (BRST-closed) and also in $\text{im } Q_1$ (BRST-exact). It is a “pure gauge” state — a boundary, in the topological language — and is also unphysical: it carries no independent information beyond χ .
- $Q_2 \Phi = 0$ and $\Phi \notin \text{im } Q_1$: Φ is BRST-closed but not BRST-exact. This is the physically meaningful case. The vector lies in the null space of Q_2 but is *not* a column of Q_1 . It is genuinely new information, not captured by any χ .

The BRST cohomology $\ker Q_2 / \text{im } Q_1$ is the space of vectors in the third case, with the identification $\Phi \sim \Phi + Q_1 \chi$ (two vectors that differ by a BRST-exact term represent the same physical state). Its dimension is:

$$\dim(\ker Q_2 / \text{im } Q_1) = \dim(\ker Q_2) - \dim(\text{im } Q_1), \quad (37.31)$$

the dimension of the null space of Q_2 minus the dimension of the column space of Q_1 . This difference counts the physical degrees of freedom: the closed states that are not exact. In the language of linear algebra, it is the dimension of the space that is annihilated by Q_2 from the left but cannot be reached by Q_1 from the right.

This is precisely the structure of the physical Hilbert space of a gauge theory. The exact states $\text{im } Q_1$ are the unphysical (gauge) degrees of freedom that can be “gauged away.” The non-exact closed states $\ker Q_2 \setminus \text{im } Q_1$ are the genuine physical excitations: transverse photons,

colour-singlet hadrons, the physical Higgs boson. The cohomology keeps the latter and discards the former, by quotienting out the exact subspace.

Confinement from BRST. An isolated quark state $|\mathbf{k}, \alpha\rangle$ with a free colour index α is not BRST-closed: the BRST charge Q generates a colour gauge rotation that acts non-trivially on the colour index. Only colour-singlet combinations are annihilated by Q :

$$\text{Meson: } |\text{meson}\rangle \propto \delta^{\alpha\beta} |q^\alpha(\mathbf{k}_1), \bar{q}_\beta(\mathbf{k}_2)\rangle, \quad (37.32)$$

$$\text{Baryon: } |\text{baryon}\rangle \propto \varepsilon_{\alpha\beta\gamma} |q^\alpha(\mathbf{k}_1), q^\beta(\mathbf{k}_2), q^\gamma(\mathbf{k}_3)\rangle. \quad (37.33)$$

These are BRST-closed because all colour indices are contracted to a singlet. BRST therefore implements confinement kinematically: the physical Hilbert space contains only colour-singlet states, not isolated quarks or gluons.

37.7 Slavnov–Taylor identities

Since the path integral measure and the gauge-fixed action are jointly BRST-invariant, the BRST current is conserved: $\langle \delta_{\text{BRST}} \mathcal{O} \rangle = 0$ for any local operator $\mathcal{O}$. These are the *Slavnov–Taylor identities* — the non-Abelian generalisations of the Ward–Takahashi identities of QED.

QED: universal electric charge. The Ward–Takahashi identity $k_\mu \Gamma^\mu(p, p+k) = S_F^{-1}(p+k) - S_F^{-1}(p)$ relates the vertex function Γ^μ to the fermion propagator S_F . A consequence is $Z_1 = Z_2$: the vertex renormalization equals the fermion wavefunction renormalization. Combined with the photon wavefunction renormalization Z_3 , this gives:

$$e_{\text{ren}} = Z_e e_{\text{bare}}, \quad Z_e = Z_1/(Z_2\sqrt{Z_3}) = 1/\sqrt{Z_3}. \quad (37.34)$$

Since Z_3 depends only on the photon propagator (not on which charged particle sits at the vertex), all charged particles — electron, muon, proton — have the same renormalized charge e . The universality of electric charge is a consequence of Ward–Takahashi, which is itself a consequence of BRST.

QCD: one coupling. The Slavnov–Taylor identities relate the renormalization of all gluon vertices:

$$\frac{Z_{g_s}}{Z_A^{3/2}} = \frac{Z_{3g}}{Z_A^{3/2}} = \frac{Z_{4g}}{Z_A^2} = \frac{Z_{\bar{\psi}g\psi}}{Z_\psi Z_A^{1/2}}, \quad (37.35)$$

where Z_A , Z_ψ , Z_{3g} , Z_{4g} , $Z_{\bar{\psi}g\psi}$ are the renormalization factors for the gluon, quark, three-gluon, four-gluon, and quark-gluon vertices respectively. All ratios are equal to the same Z_{g_s} . This means all couplings run with a *single* beta function $\beta(g_s)$: “QCD has one coupling constant” is the content of (39.7).

Standard Model: 19 parameters. The Slavnov–Taylor identities for the full Standard Model gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$ constrain all coupling renormalizations to run with three independent beta functions (one per gauge group factor). Combined with the other symmetry constraints, the independent parameters of the Standard Model reduce to exactly 19: 3 gauge couplings (g_s, g, g'), the Higgs parameters (μ, λ), the CKM matrix elements and CP-violating phase, and the Yukawa couplings. Without BRST there would be no such constraint, and renormalizability would require an infinite number of counterterms.

37.8 Ghost loops and asymptotic freedom

Ghost loops are not formal artefacts: they are required for asymptotic freedom and for the correct beta function of QCD.

The one-loop gluon self-energy $\Pi_{\mu\nu}(k)$ must be transverse: $k^\mu \Pi_{\mu\nu} = 0$. This is guaranteed by BRST (it is a consequence of the Ward identity for the gluon polarisation tensor). The three contributing diagrams are:

Loop	Contributes	Sign	Physics
Quark loop (N_f flavours)	$+\frac{1}{6}N_f$	Positive	Screening
Gluon loop ($N_c = 3$)	$-\frac{11}{6}N_c$	Negative	Antiscreening
Ghost loop	$+\frac{1}{6}N_c$	Positive	Cancels unphysical

Transversality of $\Pi_{\mu\nu}$ holds only when all three loops are included. The ghost loop precisely cancels the contribution of the unphysical longitudinal gluon polarisation in the gluon loop. Without ghosts, the gluon loop contributes an incorrect coefficient that violates transversality and gives a wrong beta function.

Summing all three contributions:

$$\beta(g_s) = -\frac{g_s^3}{16\pi^2} \left(\underbrace{\frac{11}{3}N_c}_{\text{gluon+ghost antiscreening}} - \underbrace{\frac{2}{3}N_f}_{\text{quarks screening}} \right) + O(g_s^5). \quad (37.36)$$

For $SU(3)$ ($N_c = 3$) with $N_f = 6$ quark flavours: $\beta(g_s) \propto -(11 - 4) = -7 < 0$. The coupling decreases at high energy: QCD is *asymptotically free*.

Without ghost loops, the $11/3$ coefficient would be replaced by $5/2$, and asymptotic freedom would be lost for $N_f \leq 10$ in the real world. Ghost loops are physically mandatory.

37.9 Anomalies: when BRST fails

An *anomaly* is the failure of a classical symmetry to survive quantization. For a gauge symmetry, this is catastrophic: if the path integral measure is not BRST-invariant, then $\langle \delta_{\text{BRST}} \mathcal{O} \rangle \neq 0$, gauge invariance is broken at the quantum level, negative-norm states cannot be removed from the physical spectrum, and the S-matrix is non-unitary.

The chiral (ABJ) anomaly

For a massless fermion, the axial current $j_5^\mu = \bar{\psi} \gamma^\mu \gamma^5 \psi$ is classically conserved: $\partial_\mu j_5^\mu = 0$. At one loop, a triangle diagram with one axial vertex and two vector vertices gives the Adler–Bell–Jackiw (ABJ) anomaly:

$$\partial_\mu \langle j_5^\mu \rangle = \frac{e^2}{16\pi^2} F_{\mu\nu} \tilde{F}^{\mu\nu}, \quad \tilde{F}^{\mu\nu} = \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma}. \quad (37.37)$$

The anomaly arises because the path integral measure $\mathcal{D}\psi \mathcal{D}\bar{\psi}$ is not invariant under chiral transformations $\psi \rightarrow e^{i\alpha\gamma^5} \psi$. The Jacobian of the transformation is non-trivial (Fujikawa method) and produces the $F\tilde{F}$ term. The anomaly is topological: $\int F \wedge F$ is a topological invariant related to the Atiyah–Singer index theorem.

For a *vector* current (ordinary gauge symmetry), the analogous triangle diagram must vanish for gauge invariance to be preserved at the quantum level.

Anomaly cancellation in the Standard Model

For a chiral gauge theory to be consistently quantized, the anomalies from all fermion species must cancel. The anomaly from a single left-handed Weyl fermion in representation R of the

gauge group is proportional to:

$$\mathcal{A}^{abc} = \text{Tr}_R[T^a\{T^b, T^c\}], \quad (37.38)$$

where the trace is over the representation matrices and T^a are the generators. Right-handed fermions contribute $-\mathcal{A}^{abc}$ (opposite sign, from the CPT conjugate diagram). The anomaly cancellation condition is:

$$\sum_{\text{left}} \text{Tr}_R[T^a\{T^b, T^c\}] - \sum_{\text{right}} \text{Tr}_R[T^a\{T^b, T^c\}] = 0. \quad (37.39)$$

There are four independent anomaly conditions for the Standard Model gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$:

$[U(1)_Y]^3$ **anomaly**. The condition requires:

$$\sum_{\text{left}} Y^3 - \sum_{\text{right}} Y^3 = 0. \quad (37.40)$$

Substituting the hypercharges for one generation (with colour multiplicity 3 for quarks):

$$\begin{aligned} \text{Left: } & 3 \times \left(+\frac{1}{6}\right)^3 \times 2 \text{ (doublet)} + 1 \times \left(-\frac{1}{2}\right)^3 \times 2 \text{ (doublet)} = 3 \times 2 \times \frac{1}{216} - 2 \times \frac{1}{8}, \\ \text{Right: } & 3 \times \left(+\frac{2}{3}\right)^3 + 3 \times \left(-\frac{1}{3}\right)^3 + 1 \times (-1)^3. \end{aligned}$$

Evaluating:

$$\begin{aligned} \text{Left} &= 3 \cdot 2 \cdot \frac{1}{216} + 2 \cdot \left(-\frac{1}{8}\right) = \frac{1}{36} - \frac{1}{4} = -\frac{2}{9}, \\ \text{Right} &= 3 \cdot \frac{8}{27} + 3 \cdot \left(-\frac{1}{27}\right) + (-1) = \frac{24}{27} - \frac{3}{27} - 1 = \frac{21}{27} - 1 = -\frac{2}{9}. \end{aligned}$$

Left = Right: the $[U(1)_Y]^3$ anomaly cancels.

$[SU(3)_c]^2[U(1)_Y]$ **mixed anomaly**. The condition requires $\sum_{\text{quarks}, L} Y - \sum_{\text{quarks}, R} Y = 0$:

$$2 \times \frac{1}{6} - \frac{2}{3} - \left(-\frac{1}{3}\right) = \frac{1}{3} - \frac{1}{3} = 0. \quad \checkmark \quad (37.41)$$

$[SU(2)_L]^2[U(1)_Y]$ **mixed anomaly**. Only left-handed doublets contribute (right-handed singlets have no $SU(2)_L$ coupling):

$$3 \times Y_{Q_L} + Y_{L_L} = 3 \times \frac{1}{6} + \left(-\frac{1}{2}\right) = \frac{1}{2} - \frac{1}{2} = 0. \quad \checkmark \quad (37.42)$$

Gravitational $\times [U(1)_Y]$ **anomaly**.

$$\sum_{\text{left}} Y - \sum_{\text{right}} Y = 2 \times 3 \times \frac{1}{6} + 2 \times \left(-\frac{1}{2}\right) - 3 \times \frac{2}{3} - 3 \times \left(-\frac{1}{3}\right) - (-1) = 0. \quad \checkmark \quad (37.43)$$

All anomaly conditions cancel exactly for one generation of Standard Model fermions. The cancellation is non-trivial: it requires the precise hypercharge assignments and the exact matching of quark and lepton representations. The number of colours ($N_c = 3$) enters essentially — for $N_c = 2$ or $N_c = 4$, anomalies would not cancel. This is not an accident but a consistency condition: the Standard Model is anomaly-free, and therefore consistently quantized and unitary, because of the specific fermion content.

Anomaly	Condition	Check
$[SU(3)_c]^3$	$\text{Tr}_{SU(3)}[T^a\{T^b, T^c\}] = 0$	Auto (generators traceless)
$[SU(2)_L]^3$	$\text{Tr}_{SU(2)}[\tau^a\{\tau^b, \tau^c\}] = 0$	Auto (same reason)
$[U(1)_Y]^3$	$\sum_L Y^3 = \sum_R Y^3$	Non-trivial: quarks vs leptons
$[SU(3)]^2[U(1)]$	$\sum_{q,L} Y = \sum_{q,R} Y$	Quark hypercharges
$[SU(2)]^2[U(1)]$	$\sum_{\text{doublets}} Y = 0$	Q_L vs L_L
Gravitational $\times [U(1)]$	$\sum_L Y = \sum_R Y$	Full cancellation

37.10 The Fock picture: BRST and the physical Hilbert space

The Schrödinger RNN $\dot{h}_t = -iH_{\text{gf}} h_t$ acts on the *extended* Fock space $\mathcal{F}_{\text{ext}}$, which includes the ghost sectors $\mathcal{F}_{c^a} \otimes \mathcal{F}_{\bar{c}^a}$ in addition to the physical particle sectors. Three properties ensure that this extended evolution is consistent with physical observability.

Property 1: $[H_{\text{gf}}, Q] = 0$. The gauge-fixed Hamiltonian commutes with the BRST charge. This follows from the BRST invariance of the gauge-fixed action: the BRST transformation is a symmetry of S_{gf} , so by Noether's theorem the corresponding charge Q is conserved, meaning $[H, Q] = 0$.

The operational consequence: if $h_0 \in \ker Q$ (physical initial state), then $h_t = e^{-iHt} h_0 \in \ker Q$ for all t (remains physical). The physical subspace is invariant under the Schrödinger evolution. The RNN stays within the physical Hilbert space.

Property 2: BRST-exact states decouple. For any BRST-exact state $Q|\chi\rangle \in \text{im } Q$ and any BRST-closed state $|\Phi\rangle \in \ker Q$:

$$\langle \Phi | Q | \chi \rangle = \langle Q \Phi | \chi \rangle = 0, \quad (37.44)$$

since $Q|\Phi\rangle = 0$ implies $\langle \Phi | Q^\dagger = 0$ (and Q is Hermitian up to a sign). BRST-exact states have zero inner product with all physical states: they are invisible to any physical measurement.

Property 3: ghost-number selection. The ghost number operator $N_{\text{ghost}} = \int d^3\mathbf{x} \bar{c}^a(\mathbf{x}) c^a(\mathbf{x})$ commutes with H : $[H, N_{\text{ghost}}] = 0$. Physical states have $N_{\text{ghost}} = 0$. Since the initial state h_0 has ghost number zero and this is conserved, ghost excitations never contaminate the physical sector.

The physical Hilbert space. Combining all three properties:

$$\mathcal{H}_{\text{phys}} = \frac{\ker Q \cap \{N_{\text{ghost}} = 0\}}{\text{im } Q} \subset \mathcal{F}_{\text{ext}}. \quad (37.45)$$

This is the space of BRST-closed, ghost-number-zero, non-exact states. It contains transverse photons, transverse and longitudinal ($W^\pm$, Z^0) massive gauge bosons, the physical Higgs h , colour-singlet hadrons, and massive fermions. It does not contain ghosts, unphysical polarisations, isolated quarks, or Goldstone bosons (in unitary gauge).

Extended Fock space $\mathcal{F}_{\text{ext}}$	Physical subspace $\mathcal{H}_{\text{phys}}$
Includes ghosts $c^a, \bar{c}^a$	Ghost-free ($N_{\text{ghost}} = 0$)
Includes unphysical polarisations	Transverse only (massless) or all 3 (massive)
Includes isolated coloured states	Colour singlets only
H_{gf} acts on full space	H_{SM} restricted to $\mathcal{H}_{\text{phys}}$
$[H_{\text{gf}}, Q] = 0$	Physical subspace invariant
BRST-exact states present	Decouple from all observables

The Schrödinger RNN $\dot{h}_t = -iH_{\text{SM}} h_t$, restricted to $\mathcal{H}_{\text{phys}}$, is the evolution equation for the physical universe. Ghosts, unphysical polarisations, and isolated colour charges are present in the extended Fock space used for intermediate calculations (Feynman diagrams, loop integrals), but they cancel in all physical matrix elements. They are computational scaffolding, not physical reality.

Remark. [BRST and the ML perspective: summary] The entire Faddeev–Popov–BRST construction can be summarised as a single conceptual move: trade a badly-behaved integral over an infinite-dimensional quotient space $\mathcal{A}/\mathcal{G}$ (curved geometry, non-local Jacobian, no standard Feynman rules) for a well-behaved integral over an augmented flat space $\mathcal{A} \times \mathcal{C} \times \bar{\mathcal{C}}$ (flat geometry, local Gaussian measure, standard Feynman rules), subject to the constraint that physical observables are the cohomology of the nilpotent BRST operator Q . The cost is extra degrees of freedom (ghosts); the benefit is a tractable, perturbatively computable, manifestly renormalizable theory. The BRST cohomology plays the role of a constraint layer in the computation: it projects the full extended space onto the physical subspace, exactly as a constraint in an optimisation problem projects the parameter space onto the feasible set.

37.11 Summary: from nonlinear to linear dimension reduction

The Faddeev–Popov and BRST procedures together solve a single conceptual problem: the path integral over gauge fields *overcounts* — it integrates over each physical configuration infinitely many times, once per gauge transformation. The physical configuration space is the quotient $\mathcal{A}/\mathcal{G}$ (gauge orbits), not the full field space $\mathcal{A}$. The challenge is to reduce the integration to the physical space without breaking the symmetries that make the theory tractable. The full procedure is a sequence of three steps, each translating a nonlinear problem into a linear one.

Step 1: Gauge fixing — nonlinear dimension reduction. The gauge orbit of a field configuration A_μ is the set of all configurations related to it by gauge transformations: $[A_\mu] = \{A_\mu + D_\mu \alpha : \alpha \in \mathcal{G}\}$. All elements of the same orbit describe the same physical state. The path integral naively integrates over all elements of all orbits: a massive overcounting.

Gauge fixing selects one representative per orbit by imposing a condition $G[A] = 0$ (e.g. the Lorenz condition $\partial^\mu A_\mu = 0$). This is a *nonlinear dimension reduction*: the condition $G[A] = 0$ is a surface in the infinite-dimensional field space $\mathcal{A}$, and restricting to this surface cuts out one representative from each orbit. The physical path integral should run over this surface, not over the full $\mathcal{A}$.

The problem with naive gauge fixing is the *Jacobian*: the surface $G[A] = 0$ is curved (it is not a linear subspace of $\mathcal{A}$), so changing variables from $\mathcal{A}$ to $G[A] = 0$ introduces a functional determinant $\det(\delta G/\delta \alpha)$ that depends on A_μ and cannot be ignored. If the Jacobian is dropped, the resulting path integral is inconsistent: it gives wrong answers for gauge-invariant observables.

Step 2: Faddeev-Popov — absorbing the Jacobian with ghost fields. The Faddeev-Popov trick handles the Jacobian by exponentiating it using anticommuting (Grassmann) ghost fields c^a and $\bar{c}^a$:

$$\det\left(\frac{\delta G}{\delta \alpha}\right) = \int \mathcal{D}c \mathcal{D}\bar{c} \exp\left(- \int d^4x \bar{c}^a \frac{\delta G^a}{\delta \alpha^b} c^b\right). \quad (37.46)$$

The ghost fields are not physical particles; they are an algebraic device that makes the Jacobian appear as part of the action rather than as a prefactor. The effect is to *flatten the field space*: after introducing the ghosts, the combined field space $\mathcal{A} \times \mathcal{G}_{\text{ghost}}$ (gauge fields plus ghosts) is treated as if it were a flat linear space, with the curvature of the gauge-fixing surface absorbed into the ghost action. The path integral over the flat extended space:

$$Z = \int \mathcal{D}A \mathcal{D}c \mathcal{D}\bar{c} e^{-S_{\text{YM}}[A] - S_{\text{gf}}[A] - S_{\text{ghost}}[A, c, \bar{c}]} \quad (37.47)$$

is now well-defined and tractable. The nonlinear dimension reduction (gauge fixing on a curved surface) has been traded for a linear integration over an extended flat space, at the cost of introducing ghost fields.

This is a simultaneous *dimension reduction* (gauge fixing removes the unphysical gauge directions) and *dimension augmentation* (ghost fields are added to absorb the Jacobian). The two operations cancel each other's formal effect on the partition function, leaving the correct physical answer.

Step 3: BRST — linear dimension reduction to a quotient subspace. The gauge-fixed action (Yang-Mills plus gauge-fixing plus ghosts) has a residual symmetry: the BRST transformation $\mathcal{Q}$, a nilpotent operator ($\mathcal{Q}^2 = 0$) that acts on the extended field space. The physical Hilbert space is the BRST cohomology:

$$\mathcal{H}_{\text{phys}} = \ker \mathcal{Q} / \text{im } \mathcal{Q}, \quad (37.48)$$

the quotient of the kernel of $\mathcal{Q}$ by its image. This is *linear* dimension reduction: the kernel $\ker \mathcal{Q}$ is a linear subspace (states annihilated by $\mathcal{Q}$), and the image $\text{im } \mathcal{Q}$ is a linear subspace (states of the form $\mathcal{Q}|\chi\rangle$). The quotient (37.48) identifies states that differ by a BRST-exact term, reducing the space to the physical subspace by a linear identification.

The nilpotency $\mathcal{Q}^2 = 0$ guarantees that $\text{im } \mathcal{Q} \subset \ker \mathcal{Q}$: every BRST-exact state is BRST-closed, so the quotient is well-defined. The physical states are BRST-closed but not BRST-exact: they are in the kernel but not in the image. The ghost fields, the longitudinal gauge modes, and all other unphysical degrees of freedom are either not in $\ker \mathcal{Q}$ (they are not annihilated by $\mathcal{Q}$) or are in $\text{im } \mathcal{Q}$ (they are BRST-exact and hence equivalent to zero in the cohomology). Only the transverse gauge bosons and physical matter fields survive in $\mathcal{H}_{\text{phys}}$.

The complete picture. The three steps form a coherent sequence:

Step	Operation	Type	Cost
Gauge fixing	Select one representative per orbit: $G[A] = 0$	Nonlinear dimension reduction	Jacobian $\det(\delta G/\delta \alpha)$
Faddeev-Popov	Exponentiate Jacobian as ghost action	Nonlinear $\rightarrow$ linear (flatten the space)	Ghost fields $c, \bar{c}$ added
BRST cohomology	Quotient by $\text{im } \mathcal{Q}$ in $\ker \mathcal{Q}$	Linear dimension reduction	None: linear subspace quotient

The gauge orbit structure of the path integral is a nonlinear geometry: the orbits are curved surfaces in $\mathcal{A}$, and the quotient $\mathcal{A}/\mathcal{G}$ is a curved manifold. Directly integrating over this curved

manifold is difficult. The Faddeev-Popov procedure converts the curved-manifold integration into a flat-space integration over an extended field space $\mathcal{A} \times \mathcal{G}_{\text{ghost}}$, with the curvature absorbed into the ghost action. The BRST operator then provides the final, clean, linear characterisation of the physical subspace as a cohomology. The result is that a fundamentally nonlinear geometric problem — integrating over the space of gauge orbits — is solved by a sequence of linear algebraic operations: gauge fixing (a linear constraint in the extended space), ghost fields (linear Grassmann variables), and BRST cohomology (a linear quotient).

The ML analogy. In representation learning, a similar structure arises when the data lives on a nonlinear manifold (e.g. the manifold of images of a rotating object) and one wishes to extract a linear, disentangled representation. The nonlinear manifold must first be “flattened” (by an encoder, analogous to the Faddeev-Popov flattening) and then the redundant directions (the rotational symmetry, analogous to the gauge symmetry) must be quotiented out (by identifying equivalent representations, analogous to the BRST cohomology). The ghost fields are the algebraic price of flattening a curved quotient into a linear one; in representation learning, the analogous price is the need for a carefully designed encoder that preserves the metric of the original manifold while linearising the symmetry directions.

38 Feynman Rules for the Standard Model

The generating-functional hierarchy $Z[J] \rightarrow W[J] \rightarrow \Gamma[\bar{\phi}]$ was developed for the scalar field in sections 18.5–21. The structure for the Standard Model is identical: split the action into free and interacting parts, Taylor-expand $e^{-S_{\text{int}}}$, apply Wick's theorem, track indices through contractions. The difference is the field content: multiple field types with Dirac, colour, and Lorentz indices; propagators that carry matrix structure; vertices determined by gauge couplings; and Grassmann sign rules for fermions.

This section derives the Feynman rules directly for the Standard Model, reading every propagator and vertex from the Lagrangian.

38.1 Propagators from the quadratic part

Each field type has a kinetic term in the Lagrangian. Inverting the kinetic operator gives the propagator: the Wick contraction of two free fields.

Higgs propagator. Kinetic term $\frac{1}{2}(\partial_\mu h)^2 + \frac{1}{2}m_h^2 h^2$ gives the scalar propagator:

$$\tilde{\Delta}_h(p) = \frac{i}{p^2 - m_h^2 + i\varepsilon} \quad (\text{Minkowski}). \quad (38.1)$$

One index: none. The simplest propagator in the Standard Model.

Fermion propagator. Kinetic term $\bar{\Psi}(i\not{\partial} - m)\Psi$ gives:

$$\tilde{S}_F^{AB;\alpha\beta}(p) = \frac{i(\not{p} + m)^{AB}}{p^2 - m^2 + i\varepsilon} \delta^{\alpha\beta}. \quad (38.2)$$

Two types of index: Dirac indices $A, B \in \{1, 2, 3, 4\}$ carried by the matrix $(\not{p} + m)^{AB} = (\gamma^\mu p_\mu + m)^{AB}$, and colour indices $\alpha, \beta \in \{1, 2, 3\}$ with $\delta^{\alpha\beta}$ (colour is conserved in free propagation). The fermion propagator is a 4×4 matrix in Dirac space and diagonal in colour.

Photon propagator. Maxwell kinetic term $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ with Feynman gauge-fixing $-\frac{1}{2\xi}(\partial^\mu A_\mu)^2$ ($\xi = 1$):

$$\tilde{\Delta}^{\mu\nu}(p) = \frac{-i\eta^{\mu\nu}}{p^2 + i\varepsilon}. \quad (38.3)$$

Two Lorentz indices μ, ν that contract with γ^μ, γ^ν at the QED vertices.

Gluon propagator. Identical structure, with an additional adjoint colour factor:

$$\tilde{\Delta}^{ab;\mu\nu}(p) = \frac{-i\delta^{ab}\eta^{\mu\nu}}{p^2 + i\varepsilon}, \quad (38.4)$$

with $a, b \in \{1, \dots, 8\}$ for $SU(3)_c$. The δ^{ab} ensures gluons of different adjoint colour do not mix in free propagation.

$W^\pm$ and Z^0 propagators. Massive vector bosons:

$$\tilde{\Delta}_{W,Z}^{\mu\nu}(p) = \frac{-i(\eta^{\mu\nu} - p^\mu p^\nu / M_{W,Z}^2)}{p^2 - M_{W,Z}^2 + i\varepsilon}, \quad (38.5)$$

in unitary gauge, or $-i\eta^{\mu\nu}/(p^2 - M^2 + i\varepsilon)$ in Feynman gauge (simpler for calculations, with Goldstone-boson propagators absorbing the longitudinal component).

Ghost propagator. The Faddeev–Popov ghosts c^a , $\bar{c}^a$ are anticommuting scalars in the adjoint representation:

$$\tilde{\Delta}_{\text{ghost}}^{ab}(p) = \frac{i\delta^{ab}}{p^2 + i\varepsilon}. \quad (38.6)$$

Ghosts appear only in internal lines (ghost-number conservation) and contribute (-1) per closed ghost loop (Grassmann statistics).

38.2 Vertices from the interaction Lagrangian

Each interaction term in $\mathcal{L}_{\text{SM}}$ generates a vertex. The vertex factor is read directly from the Lagrangian term, with a factor of $-i$ and a momentum-conservation delta function $(2\pi)^4\delta^4(\sum p_i)$ at each vertex.

Vertex	Lagrangian term	Vertex factor	Free indices
QED	$e\bar{\Psi}\gamma^\mu\Psi A_\mu$	$-ie(\gamma^\mu)^A_B\delta^\alpha_\beta$	μ, A, B
QCD q - g	$g_s\bar{\psi}\gamma^\mu T^a\psi G_\mu^a$	$-ig_s(\gamma^\mu)^A_B(T^a)^\alpha_\beta$	$\mu, a, A, B, \alpha, \beta$
3-gluon	$g_s f^{abc}(\partial_\mu G_\nu^a)G^{b\mu}G^{c\nu}$	$-g_s f^{abc}V^{\mu\nu\rho}(p, q, r)$	μ, ν, ρ, a, b, c
4-gluon	$g_s^2 G^4$	$-ig_s^2[\dots]_{abcd}^{\mu\nu\rho\sigma}$	4 Lorentz, 4 colour
W -fermion	$\frac{g}{\sqrt{2}}\bar{\psi}_L\gamma^\mu\psi'_L W_\mu^+$	$-i\frac{g}{\sqrt{2}}(\gamma^\mu P_L)^A_B$	μ, A, B
Z -fermion	$\frac{g}{\cos\theta_W}\bar{\psi}\gamma^\mu(g_V - g_A\gamma_5)\psi Z_\mu$	$-i\frac{g}{\cos\theta_W}(\gamma^\mu)^A_B(g_V - g_A\gamma_5)$	μ, A, B
Yukawa	$\frac{m_f}{v}\bar{\Psi}h\Psi$	$-i\frac{m_f}{v}\delta_B^A$	A, B
Ghost-gluon	$g_s f^{abc}\bar{c}^a(\partial^\mu c^b)G_\mu^c$	$-g_s f^{abc}p_c^\mu$	μ, a, b, c
Higgs self	$\frac{m_h^2}{2v}h^3 + \frac{m_h^2}{8v^2}h^4$	$-i\frac{3m_h^2}{v}, -i\frac{3m_h^2}{v^2}$	none

The three-gluon Lorentz structure: $V^{\mu\nu\rho}(p, q, r) = \eta^{\mu\nu}(p - q)^\rho + \eta^{\nu\rho}(q - r)^\mu + \eta^{\rho\mu}(r - p)^\nu$ (all momenta outgoing). The left-chirality projector: $P_L = \frac{1}{2}(1 - \gamma_5)$.

Every vertex factor is read from the Lagrangian. No guesswork, no memorisation: the Lagrangian determines the vertex, and the vertex determines the diagram.

38.3 Wick contractions with index tracking

For each term in the Taylor expansion of $e^{-S_{\text{int}}}$, Wick's theorem contracts pairs of fields using the free propagators. The Standard Model requires tracking indices through every contraction.

Fermion Wick contraction.

$$\langle\psi^{A;\alpha}(x)\bar{\psi}^{B;\beta}(y)\rangle_0 = S_F^{AB}(x - y)\delta^{\alpha\beta}. \quad (38.7)$$

Dirac indices A, B are carried by the propagator matrix S_F^{AB} ; colour indices α, β pass through as $\delta^{\alpha\beta}$. The Grassmann ordering matters: $\langle\bar{\psi}\psi\rangle = -\langle\psi\bar{\psi}\rangle$.

The fermion loop rule. A closed fermion loop at order n in the coupling gives n fermion propagators forming a cycle. The Grassmann sign rule contributes one factor of (-1) . The Dirac indices form a closed chain, producing a trace:

$$\text{fermion loop} \rightarrow (-1) \times \text{Tr}[\gamma^{\mu_1} S_F(k_1) \gamma^{\mu_2} S_F(k_2) \cdots \gamma^{\mu_n} S_F(k_n)] \times \int \frac{d^4 k}{(2\pi)^4}. \quad (38.8)$$

The trace is forced: the Dirac index that exits the last propagator must equal the index that enters the first vertex.

Gluon and photon Wick contractions.

$$\langle G^{a;\mu}(x) G^{b;\nu}(y) \rangle_0 = \delta^{ab} \Delta^{\mu\nu}(x - y). \quad (38.9)$$

Adjoint colour indices a, b pass through as δ^{ab} ; Lorentz indices μ, ν are carried by the propagator structure.

38.4 Colour factors: a worked example

The tracking of colour indices generates *colour factors*: algebraic invariants of the gauge group that multiply each diagram. We work through the quark-loop contribution to the gluon self-energy in full.

Quark loop in the gluon propagator. An external gluon $G_\mu^a(q)$ couples to a quark at vertex 1; the quark propagates around a loop; a second external gluon $G_\nu^b(q)$ couples at vertex 2. At each vertex, the QCD Lagrangian contributes $g_s(\gamma^\mu)^A_B (T^a)^\alpha_\beta$.

The colour factor: the colour indices on the two T^a matrices form a closed loop:

$$\sum_{\alpha, \beta} (T^a)^\alpha_\beta (T^b)^\beta_\alpha = \text{Tr}(T^a T^b) = \frac{1}{2} \delta^{ab}. \quad (38.10)$$

Combined with the fermion loop sign and the Dirac trace:

$$\Pi_{\mu\nu}^{ab}(q)|_{\text{quark loop}} = -g_s^2 \frac{\delta^{ab}}{2} \int \frac{d^4 k}{(2\pi)^4} \text{Tr}[\gamma_\mu S_F(k) \gamma_\nu S_F(k - q)]. \quad (38.11)$$

The result factorises: the colour factor $\text{Tr}(T^a T^b) = \frac{1}{2} \delta^{ab}$ multiplies the Dirac trace $\text{Tr}[\gamma_\mu S_F \gamma_\nu S_F]$ and the loop integral. This factorisation—colour times Dirac times kinematics—holds for all one-loop diagrams.

The Casimir invariants. The quadratic Casimirs that appear in one-loop calculations:

$$\text{Fundamental: } (T^a T^a)^\alpha_\beta = C_F \delta^\alpha_\beta, \quad C_F = \frac{N_c^2 - 1}{2N_c} = \frac{4}{3} \quad (SU(3)), \quad (38.12)$$

$$\text{Adjoint: } f^{acd} f^{bcd} = C_A \delta^{ab}, \quad C_A = N_c = 3 \quad (SU(3)). \quad (38.13)$$

C_F enters quark self-energies and vertex corrections. C_A enters gluon self-energies and the running of α_s . The ratio $C_A/C_F = 9/4$ controls the relative strength of gluon self-interaction versus quark-gluon coupling.

38.5 Ghost loops and gauge-invariance preservation

In non-Abelian gauge theories, ghost loops are required for the consistency of the perturbative expansion. The transversality of the gluon self-energy:

$$q^\mu \Pi_{\mu\nu}^{ab}(q) = 0 \quad (38.14)$$

holds only when quark loops, gluon loops, and ghost loops are all included. The ghost loop cancels the unphysical longitudinal polarisation of the gluon loop, as guaranteed by the BRST Ward identity (section on Faddeev–Popov and BRST).

The ghost-loop contribution to the gluon self-energy has colour factor $f^{acd} f^{bcd} = C_A \delta^{ab}$ and a (-1) from Grassmann statistics: the same sign rule as a fermion loop. The ghost contribution enters the one-loop beta function of QCD with the coefficient that, combined with the gluon self-interaction, produces the negative beta function (asymptotic freedom).

38.6 The complete Standard Model Feynman rules

Element	Lagrangian origin	Factor
Higgs propagator	$\frac{1}{2}(\partial h)^2 + \frac{1}{2}m_h^2 h^2$	$\frac{i}{p^2 - m_h^2 + i\varepsilon}$
Fermion propagator	$\bar{\Psi}(i\cancel{\partial} - m)\Psi$	$\frac{i(\cancel{p} + m)^{AB}}{p^2 - m^2 + i\varepsilon} \delta^{\alpha\beta}$
Photon propagator	$-\frac{1}{4}F^2$, Feynman gauge	$\frac{-i\eta^{\mu\nu}}{p^2 + i\varepsilon}$
Gluon propagator	$-\frac{1}{4}(G^a)^2$, Feynman gauge	$\frac{-i\delta^{ab}\eta^{\mu\nu}}{p^2 + i\varepsilon}$
Ghost propagator	Faddeev–Popov	$\frac{i\delta^{ab}}{p^2 + i\varepsilon}$
QED vertex	$e\bar{\Psi}\gamma^\mu\Psi A_\mu$	$-ie(\gamma^\mu)^A{}_B$
QCD q - g vertex	$g_s\bar{\psi}\gamma^\mu T^a\psi G_\mu^a$	$-ig_s(\gamma^\mu)^A{}_B(T^a)^{\alpha}{}_{\beta}$
3-gluon vertex	$g_s f^{abc}(\partial G)G^2$	$-g_s f^{abc}V^{\mu\nu\rho}$
4-gluon vertex	$g_s^2 G^4$	$-ig_s^2[\dots]$
W -fermion	$\frac{g}{\sqrt{2}}\bar{\psi}_L\gamma^\mu\psi'_L W^+$	$-i\frac{g}{\sqrt{2}}(\gamma^\mu P_L)^A{}_B$
Z -fermion	$\frac{g}{\cos\theta_W}\bar{\psi}\gamma^\mu(g_V - g_A\gamma_5)\psi Z$	$-i\frac{g}{\cos\theta_W}(\gamma^\mu)^A{}_B(g_V - g_A\gamma_5)$
Yukawa	$\frac{m_f}{v}\bar{\Psi}h\Psi$	$-i\frac{m_f}{v}\delta_B^A$
Ghost-gluon	$g_s f^{abc}\bar{c}(\partial c)G$	$-g_s f^{abc}p_c^\mu$
Higgs self	$\frac{m_h^2}{2v}h^3$	$-i\frac{3m_h^2}{v}$

Global rules:

(-1) per closed fermion or ghost loop. $\int d^4k/(2\pi)^4$ per loop momentum. $1/S$ for symmetry factor. $(2\pi)^4\delta^4(\sum p)$ at each vertex. $\text{Tr}[\dots]$ per closed fermion loop.

38.7 The Fock-Born diagnostic for Feynman diagrams

Every element of a Feynman diagram is a Fock-layer object:

The *propagator* is a Fock-layer matrix element—the two-point function $G_0^{(2)}$ of the free theory. It is not a particle travelling from x to y ; it is a component of the engine’s Green’s function.

The *vertex* is a Fock-layer coupling coefficient—the interaction term in H_{int} (or equivalently in $\mathcal{L}_{\text{int}}$). It is not an event (a particle emitted, absorbed, or scattered); it is the algebraic structure of the perturbative expansion of a matrix element.

The *internal line* is a Wick contraction between two interaction vertices. It carries an integrated loop momentum $\int d^4k$. It is not a “virtual particle”: it is a term in the cumulant expansion of $W[J] = \ln Z[J]$. Virtual particles do not appear in the Born layer. They are not detected; they are not observed; they are bookkeeping devices that organise the perturbative evaluation of Fock-layer matrix elements.

The *external line* connects the diagram to the LSZ reduction formula: amputation and on-shell evaluation convert the connected correlator into the S -matrix element $\langle f|S|i\rangle$. The squared modulus $|\langle f|S|i\rangle|^2$ is the Born probability—the cross section.

The complete pipeline:

$$\mathcal{L}_{\text{SM}} \xrightarrow{\text{Feynman rules}} G_{\text{conn}}^{(n)} \xrightarrow{\text{LSZ}} \mathcal{M} \xrightarrow{|\cdot|^2} \sigma. \quad (38.15)$$

Everything to the left of $|\cdot|^2$ is Fock-layer. The cross section σ is Born-layer. The diagram is a tool for computing the left side; the experiment measures the right side.

38.8 Worked example: $e^+e^- \rightarrow \mu^+\mu^-$ from diagram to cross section

We now compute a Born-layer observable—the cross section for $e^+e^- \rightarrow \mu^+\mu^-$ —from start to finish. This is the simplest non-trivial process that exercises the full Fock-to-Born pipeline, and it illustrates every step of the rendering process.

The calculation proceeds in five stages, each explicitly labelled as Fock-layer or Born-layer.

Stage 1: the Feynman diagram (Fock layer). At tree level ($O(e^2)$), there is one diagram: an electron and a positron annihilate into a virtual photon, which creates a muon pair. One internal photon propagator, two QED vertices.

The diagram is a Fock-layer bookkeeping device: it organises the leading-order term in the Dyson series for the S -matrix element $\langle \mu^+\mu^- | S | e^+e^- \rangle$. The “virtual photon” is not a Born-layer event; it is an internal propagator factor in a matrix element.

Stage 2: the amplitude $\mathcal{M}$ (Fock layer). Reading the diagram using the Feynman rules from the table in section 38.6:

Vertex 1 ($e^-e^+\gamma$): factor $-ie(\gamma^\mu)^A_B$, with the electron line carrying Dirac indices A, B .

Photon propagator: factor $-i\eta_{\mu\nu}/(q^2 + i\varepsilon)$, where $q = p_1 + p_2$ is the total four-momentum of the e^+e^- pair.

Vertex 2 ($\mu^+\mu^-\gamma$): factor $-ie(\gamma^\nu)^C_D$, with the muon line carrying Dirac indices C, D .

Assembling (and using momentum conservation):

$$-i\mathcal{M} = \bar{v}(p_2)(-ie\gamma^\mu)u(p_1) \frac{-i\eta_{\mu\nu}}{q^2} \bar{u}(k_1)(-ie\gamma^\nu)v(k_2), \quad (38.16)$$

where $u(p_1)$, $\bar{v}(p_2)$ are the electron and positron spinors and $\bar{u}(k_1)$, $v(k_2)$ are the muon and antimuon spinors.

Simplifying:

$$\mathcal{M} = \frac{e^2}{q^2} [\bar{v}(p_2)\gamma^\mu u(p_1)] [\bar{u}(k_1)\gamma_\mu v(k_2)]. \quad (38.17)$$

This is a complex number—a Fock-layer matrix element—that depends on the momenta and spins of all four particles. It contains phase information: the relative phases between different spin configurations can interfere. No observer measures $\mathcal{M}$ directly.

Stage 3: the squared amplitude $|\mathcal{M}|^2$ (the Born rule). The Born probability is $|\mathcal{M}|^2$: the squared modulus of the amplitude. This step is the Born rule applied at the level of scattering theory. It erases the phase information and produces a non-negative real number.

Summing over final-state spins and averaging over initial-state spins (the standard procedure for unpolarised cross sections):

$$\overline{|\mathcal{M}|^2} = \frac{1}{4} \sum_{\text{spins}} |\mathcal{M}|^2 = \frac{e^4}{q^4} L_e^{\mu\nu} L_{\mu\nu}^{(e)}, \quad (38.18)$$

where $L_e^{\mu\nu} = \frac{1}{2}\text{Tr}[(\not{p}_2 - m_e)\gamma^\mu(\not{p}_1 + m_e)\gamma^\nu]$ is the electron tensor and $L_\mu^{(e)}$ is the muon tensor (same structure with $p \rightarrow k$, $m_e \rightarrow m_\mu$). The trace is forced: summing over spins closes the Dirac index chain into a trace, exactly as in the fermion loop rule (38.8).

Evaluating the traces. Using the trace identities $\text{Tr}[\gamma^\mu\gamma^\nu] = 4\eta^{\mu\nu}$ and $\text{Tr}[\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma] = 4(\eta^{\mu\nu}\eta^{\rho\sigma} - \eta^{\mu\rho}\eta^{\nu\sigma} + \eta^{\mu\sigma}\eta^{\nu\rho})$, in the high-energy limit $s = q^2 \gg m_e^2, m_\mu^2$:

$$L_e^{\mu\nu} = 2(p_1^\mu p_2^\nu + p_1^\nu p_2^\mu - (p_1 \cdot p_2)\eta^{\mu\nu}), \quad (38.19)$$

$$L_\mu^{\mu\nu} = 2(k_1^\mu k_2^\nu + k_1^\nu k_2^\mu - (k_1 \cdot k_2)\eta^{\mu\nu}). \quad (38.20)$$

Contracting, in the centre-of-mass frame with scattering angle θ between $\mathbf{p}_1$ and $\mathbf{k}_1$:

$$\overline{|\mathcal{M}|^2} = e^4 \frac{1 + \cos^2 \theta}{2} \quad (\text{ultra-relativistic limit}). \quad (38.21)$$

The angular dependence $1 + \cos^2 \theta$ is a Born-layer prediction: it is the probability distribution for the muon emission angle, given unpolarised beams.

Stage 4: the differential cross section (Born layer). The cross section is the Born probability per unit flux, integrated over the final-state phase space:

$$d\sigma = \frac{\overline{|\mathcal{M}|^2}}{4E_1 E_2 v_{\text{rel}}} d\Pi_2, \quad (38.22)$$

where $d\Pi_2 = \frac{d^3 k_1}{(2\pi)^3 2E_{k_1}} \frac{d^3 k_2}{(2\pi)^3 2E_{k_2}} (2\pi)^4 \delta^4(q - k_1 - k_2)$ is the two-body phase space. In the CM frame:

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{4s} (1 + \cos^2 \theta), \quad \alpha = \frac{e^2}{4\pi} \approx \frac{1}{137}. \quad (38.23)$$

The total cross section, integrating over all solid angles:

$$\boxed{\sigma(e^+e^- \rightarrow \mu^+\mu^-) = \frac{4\pi\alpha^2}{3s}}. \quad (38.24)$$

This is one of the most precisely verified predictions in particle physics. At $\sqrt{s} = m_Z = 91.2 \text{ GeV}$, the cross section is $\sigma \approx 1.4 \text{ nb}$ (nanobarns)—and the measurement at LEP agrees with the SM prediction (including electroweak corrections and QCD corrections) to better than 0.1%.

Stage 5: what the detector sees (Born layer). The experimentalist at LEP does not measure $\mathcal{M}$, $|\mathcal{M}|^2$, or even σ directly. What is measured is:

Event counts: the number of $e^+e^- \rightarrow \mu^+\mu^-$ events in a given number of e^+e^- collisions, as a function of the scattering angle θ . Each event is a Born-layer rendering: the detector (a highly degenerate macroscopic observable) clicks, and the pattern of clicks is recorded.

The angular distribution: a histogram of events versus $\cos \theta$, fitted to $1 + \cos^2 \theta$. The fit confirms the vector coupling of the photon (a scalar mediator would give an isotropic distribution; a tensor mediator would give a different angular dependence).

The total cross section: the integrated event rate, divided by the luminosity. The result $\sigma = 4\pi\alpha^2/(3s)$ is compared to the SM prediction.

The full pipeline, summarised.

Stage	Object	Layer	What it is
1	Feynman diagram	Fock	Bookkeeping device for Dyson series term
2	$\mathcal{M} \in \mathbb{C}$	Fock	S -matrix element (complex, interferes)
3	$ \mathcal{M} ^2 \in \mathbb{R}_{\geq 0}$	Born rule	Phase erased; squared modulus
4	$d\sigma/d\Omega$	Born	Event rate per solid angle per flux
5	Detector clicks	Born	Macroscopic eigenvalues rendered

The Fock-to-Born transition occurs at Stage 3: taking the squared modulus. Everything before Stage 3 is internal to the engine; everything after is rendered to the observer. The squared modulus is the Born rule, applied at the level of scattering theory.

Remark. The same pipeline applies to every process at the LHC: $pp \rightarrow H + X$ (Higgs production), $pp \rightarrow t\bar{t}$ (top pair production), $pp \rightarrow jj$ (dijet production). The only changes are the diagrams (more vertices, more colour factors, more complicated phase space) and the number of Feynman diagrams at each order. The architecture—Lagrangian $\rightarrow$ Feynman rules $\rightarrow \mathcal{M} \rightarrow |\mathcal{M}|^2 \rightarrow d\sigma \rightarrow$ detector clicks—is the same. Every collider measurement is a Born-layer rendering of a Fock-layer matrix element.

38.9 Beyond tree level: what loop corrections change

The tree-level result (38.24) is the leading-order prediction. Higher orders in perturbation theory add loop corrections that modify $\mathcal{M}$ (and hence σ) systematically.

One-loop QED corrections. At order α^3 (one loop), three classes of diagrams contribute:

Vertex corrections: a virtual photon connects two points on the same fermion line, modifying the $e\psi\gamma^\mu\psi$ vertex. The corrected vertex $\Gamma^\mu(q) = \gamma^\mu F_1(q^2) + (i\sigma^{\mu\nu}q_\nu/2m)F_2(q^2)$ introduces the form factor F_2 , whose value at $q^2 = 0$ gives the anomalous magnetic moment $a_e = (g - 2)/2 = \alpha/(2\pi) + O(\alpha^2)$: Schwinger’s famous result.

Self-energy corrections: a virtual photon loop on an external fermion line modifies the propagator. These are absorbed by wave-function renormalization $Z(a)$: the external legs are renormalized, and the physical amplitude uses α_{phys} instead of α_{bare} .

Vacuum polarization: a virtual fermion loop on the internal photon line modifies the photon propagator. This introduces the running of α : $\alpha(q^2) = \alpha(0)/(1 - \Pi(q^2))$, where $\Pi(q^2)$ is the photon self-energy. At $q^2 = m_Z^2$: $\alpha(m_Z) \approx 1/128$ instead of $\alpha(0) \approx 1/137$.

What the corrections are not. The one-loop diagrams are often described as “the electron emitting and reabsorbing a virtual photon” (vertex correction) or “the photon fluctuating into an electron-positron pair” (vacuum polarization). These are Born-layer narratives applied to Fock-layer bookkeeping. The vertex correction is a one-loop integral over an internal momentum k —a term in the perturbative expansion of a matrix element. No photon is emitted; no pair is created; no fluctuation occurs. The Fock layer is one vector, rotating under e^{-iHt} . The loop integrals organise the computation of how fast it rotates.

The Born-layer observable—the anomalous magnetic moment a_e , the running coupling $\alpha(q^2)$, the corrected cross section—is what experiments measure. The Feynman diagrams are the computational tools that produce these numbers. The “virtual particles” in the diagrams are no more real than the intermediate variables in a long multiplication.

38.10 Electroweak corrections and the Z resonance

Near the Z boson mass ($\sqrt{s} \approx m_Z = 91.2 \text{ GeV}$), the process $e^+e^- \rightarrow \mu^+\mu^-$ receives a large contribution from the Z -exchange diagram: the e^+e^- pair annihilates into a virtual Z boson, which produces the muon pair. The Z propagator near resonance:

$$\frac{i}{q^2 - m_Z^2 + im_Z\Gamma_Z} \quad (\text{Breit-Wigner form}), \quad (38.25)$$

with $\Gamma_Z \approx 2.495 \text{ GeV}$. The cross section at $\sqrt{s} \approx m_Z$ is enhanced by a factor of $\sim (m_Z/\Gamma_Z)^2 \sim 1300$ relative to the pure QED prediction: the Z resonance dominates.

The Breit-Wigner peak is a Born-layer observable: the event rate as a function of $\sqrt{s}$, peaking sharply at m_Z . The peak position gives m_Z ; the peak width gives Γ_Z ; the peak height gives the Z coupling to leptons. All three are extracted by fitting the event-rate histogram to the theoretical line shape.

The number of light neutrino species N_ν is measured from the total Z width: $\Gamma_Z = \Gamma_{\text{had}} + 3\Gamma_\ell + N_\nu\Gamma_\nu$. The result $N_\nu = 2.984 \pm 0.008$ excludes a fourth light neutrino at very high significance. This is a Born-layer determination of a Fock-layer structural question (how many neutrino creation operators exist in the Fock space): the number of invisible channels in the Z decay, measured from the total event rate, determines the particle content of $\mathcal{F}_a^{\text{SM}}$.

39 Renormalization of the Standard Model: Path Integral Side

The scalar field required three engine parameters $\theta_{\phi^4}(a) = (m^2(a), \lambda(a), Z(a))$ to satisfy the rendering-stability condition $\mathcal{O}_{\text{Born}}[\theta(a)] = \mathcal{O}_{\text{phys}} + o(a/\lambda)$. The Standard Model requires 19. This section develops the rendering-stability analysis for the full SM, using the Euclidean path integral and the heat kernel.

What is new compared to the scalar field. The scalar field had one species, one symmetry ($\mathbb{Z}_2$), and three engine parameters. The Standard Model has 17 field species, three gauge symmetries ($SU(3)_c \times SU(2)_L \times U(1)_Y$), and 19 engine parameters. The three new elements are:

1. *Gauge invariance.* The engine parameters $\theta(a)$ are not arbitrary: gauge invariance enforces relations among them. On the path-integral side, these relations are encoded in the BRST symmetry and the Slavnov–Taylor identities, which reduce the number of independent counterterms from hundreds of possible dimension-4 operators to exactly 19.

2. *Multiple species with different spins.* The heat kernel computes the one-loop effective action as a sum over species, each with its own sign (+1 for bosons, −1 for fermions and ghosts) and Lorentz multiplicity. The beta functions emerge from the coefficient $\mathbf{A}_2$; the mass corrections from $\mathbf{A}_1$; the vacuum energy from $\mathbf{A}_0$.

3. *Spontaneous symmetry breaking.* The Higgs field has a nonzero vacuum expectation value $v \neq 0$, which generates masses for the W , Z , and fermions. The rendering-stability condition must hold in the broken-symmetry vacuum: the engine parameters $\theta(a)$ must be adjusted so that the physical masses m_W , m_Z , m_f (all Born-layer observables) are a -independent.

The 19 engine parameters. The SM engine at lattice spacing a has:

$$\theta_{\text{SM}}(a) = \underbrace{(g_s(a), g(a), g'(a))}_{3 \text{ gauge couplings}}, \underbrace{(y_u(a), \dots, y_\tau(a))}_{9 \text{ Yukawa couplings}}, \underbrace{\lambda(a)}_{\text{Higgs quartic}}, \underbrace{m^2(a)}_{\text{Higgs mass}}, \underbrace{V_{\text{CKM}}(a)}_{4 \text{ mixing}}, \underbrace{\theta_{\text{QCD}}(a)}_{\text{CP phase}}. \quad (39.1)$$

Each depends on the lattice spacing a . None is “bare” or “physical”—they are engine parameters at resolution a , like the weights of a neural network at a given architecture width.

The rendering-stability condition for the SM. The connected generating functional $W_{\text{SM}}[J; \theta(a), a] = \ln Z_{\text{SM}}[J; \theta(a), a]$ must satisfy:

$$W_{\text{SM}}[J; \theta(a), a] = W_{\text{SM}}^{\text{phys}}[J] + o(a/\lambda). \quad (39.2)$$

Every Born-layer prediction—every cross section, mass, mixing angle, decay width—is a functional derivative of $W_{\text{SM}}^{\text{phys}}$, and is therefore a -independent. The 19 engine parameters $\theta_{\text{SM}}(a)$ are determined by 19 Born-layer measurements (e.g., m_W , m_Z , m_H , m_t , α , α_s , G_F , mixing angles, and CP-violating phase). All other Born-layer predictions are then parameter-free.

The SM generating functional on the lattice. On a Euclidean lattice with spacing a and N sites:

$$Z_{\text{SM}}[J; \theta(a), a] = \int \prod_n [dG_n dW_n dB_n d\phi_n d\bar{\psi}_n d\psi_n d\bar{c}_n dc_n] e^{-S_{\text{SM}}[\Phi; \theta(a), a] + \sum_n J \cdot \Phi a^4}, \quad (39.3)$$

where Φ denotes all fields collectively and the integral is over $\mathbb{R}^N$ for bosons and Grassmann variables for fermions and ghosts. On the lattice: the gauge-field integral is over $SU(3) \times SU(2) \times U(1)$ link variables (compact group integration, well-defined without gauge fixing), and the fermion integral is a Grassmann Gaussian (a determinant). Everything is finite.

39.1 The generating functional on a lattice

The SM generating functional on a lattice with spacing a :

$$Z_{\text{SM}}[J; a, \theta(a)] = \int \mathcal{D}\Phi e^{-S_{\text{SM}}[\Phi; a, \theta(a)] + \int J \cdot \Phi}, \quad (39.4)$$

where Φ denotes all fields collectively ($G_\mu^a, W_\mu^i, B_\mu, \phi, \psi_f, c^a, \bar{c}^a$), J denotes all sources, and $\theta(a)$ is the set of bare parameters at resolution a .

The stability condition:

$$W_{\text{SM}}[J; a, \theta(a)] = \ln Z_{\text{SM}}[J; a, \theta(a)] = W_{\text{SM,phys}}[J] + o(a/\lambda). \quad (39.5)$$

Every Born-layer prediction (cross section, mass, mixing angle) is a derivative of $W_{\text{SM,phys}}[J]$, and must be independent of the lattice spacing a .

39.2 BRST constrains the counterterms to 19

Without gauge symmetry, any Lorentz-scalar operator of mass dimension ≤ 4 built from the SM fields could appear as an independent counterterm in the rendering-stability condition. There are hundreds of such operators. Without constraints, each would require a separate engine parameter—the theory would be non-predictive.

BRST symmetry provides the constraint. Here is how.

The problem: too many operators. Consider the gauge sector alone. The operators of dimension ≤ 4 built from A_μ^a include: $F_{\mu\nu}^a F^{a\mu\nu}$ (gauge kinetic), $\partial_\mu A_\nu^a \partial^\mu A^{a\nu}$ (non-gauge-invariant kinetic), $A_\mu^a A^{a\mu}$ (mass term), $f^{abc} A_\mu^a A_\nu^b \partial^\mu A^{c\nu}$ (cubic vertex with independent coupling), $(A_\mu^a A^{a\mu})^2$ (quartic contact with independent coupling), and more. Each could have an independent coefficient, giving dozens of parameters instead of one gauge coupling g_s .

The BRST Ward identity (Slavnov–Taylor). After gauge fixing, the action $S_{\text{SM}} + S_{\text{gf}} + S_{\text{ghost}}$ has a residual global symmetry: the BRST transformation s , which acts as:

$$sA_\mu^a = D_\mu^{ab} c^b, \quad sc^a = -\frac{1}{2} g_s f^{abc} c^b c^c, \quad s\bar{c}^a = B^a, \quad sB^a = 0, \quad s\psi = -ig_s c^a T^a \psi, \quad (39.6)$$

where c^a is the ghost, $\bar{c}^a$ the antighost, and B^a the auxiliary field. The key property: $s^2 = 0$ (nilpotency). The effective action Γ must satisfy the Slavnov–Taylor identity:

$$\mathcal{S}(\Gamma) = 0, \quad (39.7)$$

where $\mathcal{S}$ is a functional differential operator encoding the BRST symmetry. This identity constrains Γ at every order in perturbation theory and at every lattice spacing a .

How the constraints reduce the counterterms. The Slavnov–Taylor identity (39.7) enforces:

Gauge coupling universality. All vertices involving a given gauge boson must renormalise with a single coupling renormalisation. For $SU(3)_c$: the quark-gluon vertex ($g_s \bar{\psi} \gamma^\mu T^a \psi A_\mu^a$), the three-gluon vertex ($g_s f^{abc} A^a A^b \partial A^c$), and the four-gluon vertex ($g_s^2 f^{abe} f^{cde} A^a A^b A^c A^d$) all renormalise with the same Z_{g_s} . Without BRST: three independent couplings. With BRST: one.

Transversality of gauge-boson self-energy. $q^\mu \Pi_{\mu\nu}^{ab}(q) = 0$. This forbids the gauge-boson mass counterterm $\delta M^2 A_\mu A^\mu$: gauge bosons remain massless (before spontaneous breaking) at every lattice spacing. One fewer parameter.

Ghost-gauge coupling relation. The ghost-gluon vertex renormalisation is not independent: it is determined by Z_{g_s} and the ghost wave-function renormalisation. No new parameter.

Elimination of non-gauge-invariant operators. Operators like $\partial_\mu A_\nu^a \partial^\mu A^{a\nu}$ (which differ from $F_{\mu\nu} F^{\mu\nu}$ by gauge-variant terms) are forbidden as counterterms. BRST enforces that only the gauge-invariant combination $F_{\mu\nu} F^{\mu\nu}$ appears.

The net result: the hundreds of possible dimension- ≤ 4 operators collapse to the 19 independent terms of the SM Lagrangian. Each corresponds to one engine parameter in $\theta_{\text{SM}}(a)$.

In ML terms. BRST is a weight-sharing constraint: it ties together parameters that naïvely look independent, exactly as a convolutional layer shares weights across spatial positions. The “gauge coupling universality” is the statement that the same coupling appears in every vertex involving a given gauge boson—a severe weight-sharing rule that reduces the parameter count from hundreds to 19.

39.3 The one-loop effective action: sum over species

The one-loop effective action (section 23.11):

$$\Gamma_{\text{SM}}^{(1)} = -\frac{1}{2} \int_0^\infty \frac{ds}{s} \text{Tr} e^{-s D_{\text{SM}}}. \quad (39.8)$$

The operator D_{SM} is block-diagonal: one block per field species. The trace decomposes as a sum:

$$\text{Tr} e^{-s D_{\text{SM}}} = \sum_{\text{species } i} \eta_i \text{Tr}_i e^{-s D_i}, \quad (39.9)$$

where $\eta_i = +1$ for bosons and $\eta_i = -1$ for fermions and ghosts (the Grassmann sign). Each species contributes its own Seeley–DeWitt coefficients $a_{n,i}(x)$, and the total is the sum:

$$\mathbf{A}_n = \int d^4x \sum_i \eta_i \text{tr}_i a_{n,i}(x). \quad (39.10)$$

The divergences of the SM are encoded in $\mathbf{A}_0$, $\mathbf{A}_1$, and $\mathbf{A}_2$.

39.4 The rendering-stability workflow for the Standard Model

The three-step workflow of Section 21.9, developed for the scalar field, applies to the SM with the following upgrades.

Step 1: compute Born-layer observables $\mathcal{O}[\theta_{\text{SM}}(a), a]$. Fix a and $\theta_{\text{SM}}(a)$. Compute the Euclidean correlators by Feynman diagrams on the lattice (or by Monte Carlo sampling). Extract the Born-layer observables:

Born-layer observable	Fock-layer source	Bridge
m_W, m_Z, m_H, m_f	Poles of propagators $\Gamma_i^{(2)}$	Spectral decomposition
α, α_s, G_F	Vertex functions $\Gamma^{(3)}, \Gamma^{(4)}$	Renormalization conditions
V_{ij}^{CKM}	Flavour-changing vertices	Amplitude ratios
Cross sections σ	Amputated $G_{\text{conn}}^{(n)}$	LSZ + $ \cdot ^2$

Each observable depends on $\theta_{\text{SM}}(a)$ and a .

Step 2: impose rendering stability. Choose $\theta_{\text{SM}}(a)$ so that 19 Born-layer observables are a -independent:

$$\mathcal{O}_k[\theta_{\text{SM}}(a), a] = \mathcal{O}_{k,\text{phys}} + o(a/\lambda), \quad k = 1, \dots, 19. \quad (39.11)$$

These 19 equations determine the 19 engine parameters as functions of a and the 19 physical observables. The Seeley–DeWitt coefficients $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{A}_2$ tell us how $\theta_{\text{SM}}(a)$ must run:

$\mathbf{A}_0$ (vacuum energy): absorbed into $E_0(a)$.

$\mathbf{A}_1$ (quadratic in $1/a$): determines how $m^2(a)$ runs—the Higgs mass parameter scales as $1/a^2$.

$\mathbf{A}_2$ (logarithmic in $1/a$): determines the beta functions of all gauge couplings, Yukawa couplings, and the Higgs quartic.

Step 3: verify that all other observables are automatically a -independent. This is the content of the SM’s renormalisability, proved by ’t Hooft and Veltman (1971–72). Once 19 rendering conditions are imposed, every $\Gamma^{(n)}$ with $n > 4$ is finite and a -independent at each loop order. The proof uses BRST: the Slavnov–Taylor identity (39.7) guarantees that no new a -dependent structures appear beyond those already absorbed by the 19 parameters.

The SM workflow, compared to scalar field.

	Scalar ϕ^4	Standard Model
Engine parameters	$\theta(a)$: 3 numbers	$\theta_{\text{SM}}(a)$: 19 numbers
Rendering conditions	3 (mass, coupling, wave fn.)	19 (masses, couplings, mixing angles)
Symmetry constraint	$\mathbb{Z}_2$ (mild)	$SU(3) \times SU(2) \times U(1)$ BRST (reduces hundreds to 19)
$\mathbf{A}_1$ correction	$\delta m^2 \propto \lambda/a^2$	$\delta m_H^2 \propto (6\lambda + 6y_t^2 - \frac{9}{4}g^2 - \frac{3}{4}g'^2)/a^2$
$\mathbf{A}_2$ correction	$\beta_\lambda = 3\lambda^2/(16\pi^2)$	$\beta_{g_s}, \beta_g, \beta_{g'}, \beta_{y_f}, \beta_\lambda$ (all computed)
Proof of renorm.	Power counting	’t Hooft–Veltman + BRST

39.5 Conflations specific to the gauge theory

The gauge structure of the SM introduces new opportunities for conflation, beyond those catalogued in Section ??.

“The Higgs gives mass to particles.”

Common narration: “The Higgs field permeates all of space, and particles acquire mass by interacting with it. Heavier particles interact more strongly with the Higgs.”

What actually happens: The Higgs field has a nonzero vacuum expectation value $v = \langle \hat{\phi} \rangle \neq 0$ in the ground state $|\Omega\rangle$. The Yukawa coupling $y_f \bar{\psi} \phi \psi$, evaluated at $\phi = v + h$ (where h is the fluctuation), gives a fermion mass term $m_f = y_f v / \sqrt{2}$. The mass is a property of the vacuum state (a Fock-layer datum: v is the minimum of the effective potential), not a “drag” or “interaction” that happens to particles as they “move through” the field. The fermion mass m_f is a Born-layer observable: what the spectrometer reports. The Yukawa coupling $y_f(a)$ is an engine parameter. The VEV v is a Fock-layer quantity (the expectation value of $\hat{\phi}$ in $|\Omega\rangle$).

“Ghosts are needed for unitarity.”

Common narration: “Without ghost fields, the S -matrix violates unitarity. Ghosts are unphysical particles with the wrong spin-statistics, but their contribution cancels the unphysical longitudinal and timelike gluon polarisations, restoring unitarity.”

What actually happens: In the Hamiltonian formulation (Coulomb gauge), the Fock space contains only the 2 physical transverse polarisations per gauge boson. No ghost fields appear. The S -matrix is manifestly unitary on this physical Fock space. Ghost fields arise only in

the path-integral formulation with covariant gauge fixing, where all 4 polarisations of A_μ^a are integrated over. The Faddeev–Popov determinant (rewritten as a ghost integral) subtracts the overcounting. Ghosts are path-integral bookkeeping: they enforce the Gauss law constraint that is built into the Hamiltonian from the start. No ghost has ever been detected, and none ever will—they are not Born-layer objects.

“The gauge boson propagator is a force carrier.”

What actually happens: The t -channel propagator $iD_{\mu\nu}(q)/(q^2 - M^2)$ in a scattering diagram is a Fock-layer matrix element—the Green’s function of the gauge-field equation of motion. Its Fourier transform gives the Yukawa or Coulomb potential, which encodes how the Hamiltonian couples two-particle states. No gauge boson “travels” from one particle to another. The Born-layer observable is $d\sigma/dt$: the angular distribution of scattered particles, measured by counting detector clicks.

“Renormalization removes infinities.”

Nothing is removed. The engine parameters $\theta_{\text{SM}}(a)$ are finite numbers at every finite a . The integrals on the lattice are finite sums. The Born-layer observables are finite predictions. Rendering stability requires $\theta_{\text{SM}}(a)$ to depend on a so that the Born outputs do not. The word “infinity” arises only in the continuum extrapolation $a \rightarrow 0$, where the engine parameters $\theta_{\text{SM}}(a)$ grow without bound. On the lattice, nothing is infinite, nothing is removed, and nothing is cancelled.

39.6 $\mathbf{A}_0$: vacuum energy and the cosmological constant

The $n = 0$ coefficient counts degrees of freedom:

$$\mathbf{A}_0 = \int d^4x \sum_i \eta_i n_i, \quad (39.12)$$

where n_i is the number of field components (4 for a gauge boson, 4 for a Dirac fermion, 1 for a real scalar, 1 for a ghost).

The integral $\int_0^\infty ds s^{-3} \mathbf{A}_0$ produces a quartic divergence $\sim \Lambda^4$: the vacuum energy. Each bosonic degree of freedom contributes $+\Lambda^4/(16\pi^2)$; each fermionic degree of freedom contributes $-\Lambda^4/(16\pi^2)$. The SM has 28 bosonic and 90 fermionic helicity states; the net vacuum energy is $\sim -62\Lambda^4/(16\pi^2)$, which for $\Lambda \sim M_P$ exceeds the observed cosmological constant by ~ 120 orders of magnitude.

In the lattice-first viewpoint, $\mathbf{A}_0$ is absorbed into the bare vacuum energy $E_0(a)$. The mismatch between the natural scale $1/a^4$ and the observed value $\Lambda_{\text{obs}} \sim (10^{-3} \text{ eV})^4$ is the cosmological constant problem: the most extreme fine-tuning required by the stability condition.

39.7 $\mathbf{A}_1$: mass renormalization and the hierarchy problem

The $n = 1$ coefficient involves $\text{tr } E_i$, where E_i is the endomorphism (“potential” term) of the operator D_i . For the Higgs field, E contains $m_h^2 + \lambda\bar{\phi}^2$; for fermions, E contains m_f^2 ; for gauge bosons, $E = 0$ (in the massless limit).

The integral $\int_0^\infty ds s^{-2} \mathbf{A}_1$ produces a quadratic divergence $\sim \Lambda^2$. For the Higgs mass:

$$\delta m_h^2 \propto \frac{1}{16\pi^2} \frac{1}{a^2} (6\lambda + 6y_t^2 - \frac{9}{4}g^2 - \frac{3}{4}g'^2 + \dots), \quad (39.13)$$

where the contributions come from: Higgs self-coupling ($+6\lambda$), top quark loop ($+6y_t^2$, sign from fermion), gauge boson loops ($-\frac{9}{4}g^2 - \frac{3}{4}g'^2$).

The hierarchy problem: $m_{\text{phys}} \approx 125 \text{ GeV}$ requires $m^2(a)$ to cancel the $1/a^2$ correction at each lattice spacing a . If $a \sim \ell_P$: the bare mass must be tuned to ~ 34 decimal places to achieve $m_{\text{phys}}/M_P \sim 10^{-17}$. The stability condition does not forbid this—it simply requires $m^2(a)$ to be whatever function of a produces the correct m_{phys} . The “problem” is aesthetic, not logical: the required $m^2(a)$ is an extremely rapidly varying function of a .

39.8 $\mathbf{A}_2$: the beta functions of the Standard Model

This is the payoff of the heat kernel for the Standard Model.

The $n = 2$ coefficient contains $\text{tr}(\frac{1}{2}E^2 + \frac{1}{12}\Omega_{\mu\nu}^2)$. The $\Omega_{\mu\nu}^2$ term renormalises the gauge kinetic terms and determines the gauge beta functions; the E^2 term renormalises the Higgs quartic and Yukawa couplings.

For a field in representation R of a gauge group with generators T_R^a ($\text{tr}(T^a T^b) = T(R)\delta^{ab}$):

$$\text{tr}_R \Omega_{\mu\nu}^2 = g^2 T(R) F_{\mu\nu}^a F^{a\mu\nu}. \quad (39.14)$$

Each field species contributes with sign η_i and Lorentz multiplicity $n_i^{(\mu)}$. The coefficient of $F_{\mu\nu}^a F^{a\mu\nu}$ in $\mathbf{A}_2$ is:

$$\mathbf{A}_2|_{F^2} = \frac{g^2}{12} \left[\sum_{\text{bosons}} n_i^{(\mu)} T_i - \sum_{\text{fermions}} n_i^{(\mu)} T_i - \sum_{\text{ghosts}} T_i \right] \int d^4x F^a F^a. \quad (39.15)$$

$SU(3)_c$: the QCD beta function.

Field	Rep	$n^{(\mu)}$	$T(R)$	Contrib.	Sign
Gluon G_μ^a	Adj. 8	4	3	+12	Boson
Ghost c^a	Adj. 8	1	3	−3	Ghost
$Q_{f,L}$ ($3_{\text{gen}} \times 2$)	Fund. 3	2	$\frac{1}{2}$	−6	Fermion
$u_{f,R}$ (3_{gen})	Fund. 3	2	$\frac{1}{2}$	−3	Fermion
$d_{f,R}$ (3_{gen})	Fund. 3	2	$\frac{1}{2}$	−3	Fermion
Total				−3	

With the standard normalisation:

$$\beta_{g_s} = -\frac{g_s^3}{16\pi^2} \left(\frac{11}{3} \cdot 3 - \frac{2}{3} \cdot 6 \right) = -\frac{7g_s^3}{16\pi^2}. \quad (39.16)$$

The negative sign is asymptotic freedom: the strong coupling decreases at short distances. This single number $\beta_1 = -7$ determines the running of α_s from $\alpha_s(m_Z) \approx 0.118$ down to perturbative values at high energies.

$SU(2)_L$: the weak beta function.

Field	Rep	$n^{(\mu)}$	$T(R)$	Contrib.
W_μ^i	Adj. 3	4	2	+8
Ghost ω^i	Adj. 3	1	2	−2
Higgs ϕ	Fund. 2	1	$\frac{1}{2}$	$+\frac{1}{2}$
$Q_{f,L}$ ($3_{\text{gen}} \times 3_c$)	Fund. 2	2	$\frac{1}{2}$	−9
$L_{f,L}$ (3_{gen})	Fund. 2	2	$\frac{1}{2}$	−3
Total				$-\frac{11}{2}$

$$\boxed{\beta_g = -\frac{g^3}{16\pi^2} \left(\frac{22}{3} - \frac{2}{3} \cdot 12 - \frac{1}{6} \right) = -\frac{19g^3}{96\pi^2}.} \quad (39.17)$$

Asymptotically free, though less strongly than QCD.

$U(1)_Y$: the hypercharge beta function. $U(1)$ has no self-interaction (no three-boson or four-boson vertex), so there is no antiscreening contribution. Only fermion and scalar loops contribute, all screening:

$$\boxed{\beta_{g'} = +\frac{41g'^3}{96\pi^2}.} \quad (39.18)$$

The positive sign means g' increases at short distances: the hypercharge coupling is not asymptotically free. This is the same behaviour as QED ($\beta_e > 0$), for the same reason: charged loops screen the bare charge.

39.9 Screening versus antiscreening

The QCD beta function decomposes into physically distinct contributions:

Source	Coeff.	Effect	Mechanism
Gluon loops	$-\frac{10}{3}N_c$	Antiscreening	Virtual gluons are paramagnetic: they align with the colour field, amplifying it at long distance
Ghost loops	$+\frac{1}{3}N_c$	Screening	Cancels unphysical longitudinal/timelike gluon polarisations
Quark loops	$+\frac{2}{3}N_f$	Screening	Virtual $q\bar{q}$ pairs are diamagnetic: they oppose the colour field (as in QED)
Net ($N_c=3, N_f=6$)	−7	Antiscreening	Asymptotic freedom: $\alpha_s \rightarrow 0$ at short distance

The gluon self-interaction—the commutator $[G_\mu, G_\nu]$ in the Yang–Mills field strength, forced by the non-commutativity of $SU(3)$ —produces antiscreening. This dominates the screening from quark loops for $N_f < 11N_c/2 = 33/2$: the SM has $N_f = 6$, safely below the threshold. Without ghost loops, the gluon contribution would include unphysical polarisations and give the wrong coefficient.

The Abelian group $U(1)_Y$ has no gluon self-interaction (no $[B_\mu, B_\nu]$ term), hence no antiscreening, hence $\beta_{g'} > 0$. The difference between asymptotic freedom (non-Abelian) and screening (Abelian) is the self-interaction of the gauge field—a consequence of the gauge group’s non-commutativity.

39.10 The running of α_s : a Born-layer observable

The QCD beta function $\beta_{g_s} = -7g_s^3/(16\pi^2)$ determines how the strong coupling $\alpha_s = g_s^2/(4\pi)$ changes with the momentum scale μ at which it is measured. Integrating the one-loop RG equation $\mu d\alpha_s/d\mu = -2b_3\alpha_s^2/(4\pi)$ with $b_3 = 7$:

$$\alpha_s(\mu) = \frac{\alpha_s(\mu_0)}{1 + \frac{7\alpha_s(\mu_0)}{2\pi} \ln \frac{\mu}{\mu_0}}. \quad (39.19)$$

With the measured value $\alpha_s(m_Z) = 0.1179 \pm 0.0009$ at $\mu_0 = m_Z = 91.2 \text{ GeV}$:

μ	1.5 GeV	5 GeV	91.2 GeV	1 TeV
$\alpha_s(\mu)$	~ 0.35	~ 0.21	0.118	~ 0.088
Regime	<i>Non-perturbative</i>	<i>Borderline</i>	<i>Perturbative</i>	<i>Weakly coupled</i>

At low energies ($\mu \lesssim 1 \text{ GeV}$): $\alpha_s \sim 1$ and perturbation theory fails. The quarks confine into hadrons, the coupling is strong, and the physics must be computed non-perturbatively (lattice QCD). At high energies ($\mu \gg m_Z$): $\alpha_s \ll 1$ and perturbation theory works beautifully. This is asymptotic freedom: the engine runs weakly at short distances.

Born-layer evidence. The running of α_s is a Born-layer observable: it is measured from event rates and kinematic distributions at different energies. Each measurement is a Born probability (a cross section or a decay rate) evaluated at a specific momentum transfer μ .

The evidence comes from a remarkable variety of processes:

τ decay ($\mu \sim m_\tau = 1.78 \text{ GeV}$): the ratio $R_\tau = \Gamma(\tau \rightarrow \text{hadrons})/\Gamma(\tau \rightarrow e\bar{\nu}\nu)$ is sensitive to $\alpha_s(m_\tau)$.

Jet rates in e^+e^- ($\mu \sim \sqrt{s}$): the fraction of events with 3 jets (requiring a hard gluon emission) is proportional to α_s .

Deep inelastic scattering ($\mu \sim Q$, the momentum transfer): the scaling violations of structure functions (the Q -dependence of $F_2(x, Q^2)$) are governed by the DGLAP equations, whose splitting functions are proportional to $\alpha_s(Q)$.

Higgs production ($\mu \sim m_H$): the dominant production mechanism at the LHC ($gg \rightarrow H$ via a top loop) is proportional to α_s^2 .

All measurements agree with the one-loop running (39.19) to within experimental uncertainties. The world average $\alpha_s(m_Z) = 0.1179 \pm 0.0009$ is determined from a global fit to all these data simultaneously.

Λ_{QCD} : the dimensional transmutation. The running coupling can be rewritten as:

$$\alpha_s(\mu) = \frac{2\pi}{7 \ln(\mu/\Lambda_{\text{QCD}})}, \quad \Lambda_{\text{QCD}} \approx 200\text{--}300 \text{ MeV}. \quad (39.20)$$

The dimensionless coupling α_s has been traded for a dimensionful scale Λ_{QCD} . This is *dimensional transmutation*: a dimensionless parameter of the classical Lagrangian (g_s) has been converted, by quantum effects (the running), into a mass scale (Λ_{QCD}) that sets the scale of confinement, the proton mass, and the entire hadronic spectrum.

The proton mass $m_p \approx 938 \text{ MeV}$ is mostly *not* from the quark masses ($m_u + m_d + m_u \sim 10 \text{ MeV}$, less than 2% of m_p). It is from Λ_{QCD} : the energy of the gluon field and the kinetic energy of the confined quarks. The proton mass is a Born-layer observable whose origin is dimensional transmutation—a Fock-layer quantum effect.

39.11 Gauge coupling unification

The three gauge couplings $\alpha_1, \alpha_2, \alpha_3$ run with energy according to their respective beta functions. With the SM particle content:

$$\alpha_1^{-1}(\mu) = \alpha_1^{-1}(m_Z) - \frac{41}{20\pi} \ln \frac{\mu}{m_Z}, \quad (39.21)$$

$$\alpha_2^{-1}(\mu) = \alpha_2^{-1}(m_Z) + \frac{19}{12\pi} \ln \frac{\mu}{m_Z}, \quad (39.22)$$

$$\alpha_3^{-1}(\mu) = \alpha_3^{-1}(m_Z) + \frac{7}{2\pi} \ln \frac{\mu}{m_Z}, \quad (39.23)$$

using the GUT normalisation $\alpha_1 = (5/3)\alpha_Y$ for $U(1)_Y$ (the factor 5/3 is chosen so that the three generators have the same normalisation when embedded in $SU(5)$).

The SM does not unify. With the SM particle content, the three lines $\alpha_i^{-1}(\mu)$ do *not* meet at a single point: they form a triangle at $\mu \sim 10^{14}$ – 10^{16} GeV, missing unification by several percent. This is a quantitative statement about the Born-layer running: the measured couplings at m_Z , evolved upward using the SM beta functions, do not converge.

The MSSM improves the convergence. With the MSSM particle content (all superpartner thresholds at $\mu \sim 1$ TeV), the modified beta function coefficients $(b_1, b_2, b_3) = (-33/5, -1, 3)$ cause the three coupling lines to meet at a single point:

$$\alpha_{\text{GUT}}^{-1} \approx 24, \quad M_{\text{GUT}} \approx 2 \times 10^{16} \text{ GeV}. \quad (39.24)$$

The convergence is quantitatively precise: the three lines meet to within $\sim 1\%$ at M_{GUT} . This is often cited as evidence for both supersymmetry and grand unification.

The Fock-Born reading. Gauge coupling unification is a Born-layer prediction: it concerns the values of $\alpha_i(\mu)$ at very high energies, extrapolated from low-energy measurements using the beta functions. The extrapolation is a Fock-layer computation (integrating the RG equations). The convergence (or lack thereof) is a statement about the Fock-layer engine: does the engine's parameter space have a point where all three couplings are equal?

In the lattice-first viewpoint: the three couplings $g_s(a), g(a), g'(a)$ are bare parameters of the engine at lattice spacing a . If the engine has a lattice spacing $a_{\text{GUT}} = \pi/M_{\text{GUT}}$ at which all three bare couplings are equal— $g_s(a_{\text{GUT}}) = g(a_{\text{GUT}}) = g'(a_{\text{GUT}}) = g_{\text{GUT}}$ —then the three couplings at $a \gg a_{\text{GUT}}$ (i.e., at low energies) will be different, split by the different beta functions. Grand unification is the hypothesis that the engine has this simplified structure at very short distances.

Whether this hypothesis is correct is an open Born-layer question: it will be tested if proton decay (predicted by grand unified theories with a rate $\tau_p \propto M_{\text{GUT}}^4/m_p^5$) is observed. Current bounds: $\tau_p > 10^{34}$ years (Super-Kamiokande), pushing $M_{\text{GUT}} > 10^{16}$ GeV.

39.12 Born-layer tests of the Standard Model

The precision of the Standard Model's Born-layer predictions is extraordinary. We catalogue the most precise tests:

Observable	Measurement	SM prediction	Agreement
$(g-2)_e/2$	$1159652180.73(28) \times 10^{-12}$	Requires α as input	12 significant figures
m_W	$80.3692 \pm 0.0133 \text{ GeV}$	$80.357 \pm 0.006 \text{ GeV}$	0.02%
$\sin^2 \theta_W$	0.23153 ± 0.00004	0.23152 ± 0.00005	0.004%
R_ℓ (LEP)	20.767 ± 0.025	20.744 ± 0.010	0.1%
$\alpha_s(m_Z)$	0.1179 ± 0.0009	(input)	—
$\text{BR}(B \rightarrow X_s \gamma)$	$(3.32 \pm 0.15) \times 10^{-4}$	$(3.36 \pm 0.23) \times 10^{-4}$	1%

Each row is a test of the rendering-stability condition: the Fock-layer engine at lattice spacing a (with 19 parameters fixed by 19 measurements) predicts the Born-layer number in the right column. The agreement, ranging from 0.004% to 1%, confirms that the engine produces an a -independent rendering to extraordinary precision.

The anomalous magnetic moment $(g-2)_e$ deserves special comment. The SM prediction involves diagrams up to fifth order in α (containing up to 12672 Feynman diagrams at fifth order), numerical integration over 10-dimensional parameter spaces, and hadronic and electroweak corrections. The agreement with experiment to 12 significant figures is the most precise test of any physical theory. It confirms that the Fock-layer engine—the Hamiltonian $H_{\text{SM}}(a)$ with its 19 parameters—produces a Born-layer rendering that matches reality to one part in 10^{12} .

39.13 The 19 parameters of the Standard Model

Parameter	Count	Divergence sorbed	ab- Experimental input
g_s	1	$\ln(1/a) \cdot G^2$	$\alpha_s(m_Z)$
g	1	$\ln(1/a) \cdot W^2$	G_F
g'	1	$\ln(1/a) \cdot B^2$	α_{em}
λ	1	$\ln(1/a) \cdot (\phi^\dagger \phi)^2$	m_H
μ^2	1	$(1/a^2) \cdot \phi^\dagger \phi$	v (or m_Z)
y_u, y_c, y_t	3	$\ln(1/a) \cdot \bar{Q} \tilde{\phi} u_R$	m_u, m_c, m_t
y_d, y_s, y_b	3	$\ln(1/a) \cdot \bar{Q} \phi d_R$	m_d, m_s, m_b
y_e, y_μ, y_τ	3	$\ln(1/a) \cdot \bar{L} \phi e_R$	m_e, m_μ, m_τ
$\theta_{12}, \theta_{13}, \theta_{23}, \delta$	4	$\ln(1/a)$ (Yukawa mixing)	CKM measurements
θ_{QCD}	1	Topological	$< 10^{-10}$ (strong CP)
Total	19		

The Standard Model is renormalisable: 19 bare parameters $\theta(a)$, three Seeley–DeWitt coefficients ($\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2$), all divergences absorbed by BRST-constrained parameter shifts. Every Born-layer prediction is independent of the lattice spacing a , to accuracy $o(a/\lambda)$.

Each row in the table is one axis of the rendering-stability condition: the left column is a Fock-layer parameter of the engine at resolution a ; the right column is the Born-layer measurement that fixes it. The engine has 19 dials; the screen has 19 numbers; turning the dials to match the numbers is renormalization.

39.14 Dimensional regularization as a computational tool

In practice, perturbative calculations in the SM use *dimensional regularization*: replace $d = 4$ with $d = 4 - 2\varepsilon$ and evaluate all momentum integrals in d dimensions. The divergences appear as poles in $1/\varepsilon$:

$$\int \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2 + M^2)^2} = \frac{1}{16\pi^2} \left(\frac{1}{\varepsilon} - \gamma_E + \ln \frac{4\pi\mu^2}{M^2} + O(\varepsilon) \right). \quad (39.25)$$

The $\overline{\text{MS}}$ scheme absorbs $1/\varepsilon - \gamma_E + \ln 4\pi$ into the bare parameters, leaving the finite part depending on $\ln(\mu^2/M^2)$.

The correspondence with the lattice:

Lattice (a)	Dimensional reg. (ε)
$\ln(1/a)$	$1/\varepsilon$
$1/a$ (lattice resolution)	μ (renormalization scale)
$\theta(a)$ (bare parameters)	$\theta(\mu)$ (running parameters)
$o(a/\lambda)$ stability	$\overline{\text{MS}}$ subtraction

Dimensional regularization is a computational tool, not a physical framework. It preserves gauge invariance automatically (because it respects the algebraic structure of the gauge group in d dimensions) and simplifies multi-loop calculations enormously. This is why Veltman and 't Hooft's proof of the renormalisability of Yang–Mills theory (1971–72), which introduced dimensional regularization, was a breakthrough: it provided a regularization that did not break the gauge invariance it was trying to preserve.

In the lattice-first viewpoint, dimensional regularization is a shortcut for computing the functions $\theta(a)$. The physical framework is the lattice $H(a)$ on $\mathcal{F}_a$ with the stability condition. Dimensional regularization computes the same β functions, the same anomalous dimensions, the same physical predictions—but it does so by a different bookkeeping method, one that never constructs the lattice explicitly.

39.15 The Fock-Born summary

The Standard Model as a stably finitised theory:

The Fock layer is a lattice with spacing a , Hamiltonian $H_{\text{SM}}(a)$, and 19 bare parameters $\theta(a)$. The Seeley–DeWitt coefficients $\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2$ determine the a -dependence of the one-loop effective action. The three gauge beta functions (39.16–39.18) determine how the gauge couplings run with a .

The Born layer renders 19 physical quantities (masses, couplings, mixing angles) that fix the bare parameters. Every other Born-layer prediction—every cross section at the LHC, every branching ratio at Belle II, every anomalous magnetic moment—is a derivative of $W_{\text{SM,phys}}[J]$ and is a -independent to $o(a/\lambda)$.

The SM is the most precisely tested stably finitised theory in physics. The anomalous magnetic moment of the electron $(g - 2)_e$ is measured to 0.28 parts per trillion and agrees with the SM prediction to this precision: 12 significant figures of agreement between the engine's output and the Born-layer measurement. This agreement is the rendering-stability condition verified to extraordinary accuracy.

40 Renormalization of the Standard Model: the Hamiltonian Side

The path-integral renormalization of the Standard Model (Section 39) computed the one-loop structure from the Seeley–DeWitt coefficients $\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2$, summing over all field species including Faddeev–Popov ghosts. The rendering-stability condition $W_{\text{SM}}[J; \theta(a), a] = W_{\text{SM}}^{\text{phys}}[J] + o(a/\lambda)$ determined the 19 engine parameters $\theta_{\text{SM}}(a)$ as functions of a .

This section develops the same physics from the Hamiltonian side: the Fock space $\mathcal{F}_a^{\text{SM,phys}}$ on a spatial torus $\mathbb{T}^3$ with lattice spacing a , the Hamiltonian $H_{\text{SM}}(a)$, and the Feshbach–Löwdin mode elimination that derives how $\theta_{\text{SM}}(a)$ runs with a .

The most striking difference: in the Hamiltonian approach, ghost fields never appear. Gauge invariance is enforced by the Gauss law constraint on the physical Hilbert space, not by the Faddeev–Popov procedure on the path-integral measure. The physical Fock space contains only the degrees of freedom that can produce Born-layer events: 2 transverse polarisations per gauge boson, no ghosts, no unphysical modes.

The beta functions, the running of all 19 engine parameters, and every Born-layer prediction agree exactly with the path-integral side. The decomposition into channels differs; the totals are identical.

40.1 Gauge invariance without ghosts: the Gauss law

In the canonical formulation, the gauge field A_μ^a has four components per colour index. The three spatial components A_i^a ($i = 1, 2, 3$) are dynamical: their conjugate momenta are $\Pi^{ai} = -E^{ai}$ (the colour-electric field). The temporal component A_0^a has no conjugate momentum: $\partial\mathcal{L}/\partial\dot{A}_0^a = 0$. Its Euler–Lagrange equation is not a dynamical equation but a *constraint*—the Gauss law:

$$\hat{G}^a(\mathbf{x}) \equiv D_i^{ab} \hat{E}^{bi}(\mathbf{x}) - g_s \hat{\rho}^a(\mathbf{x}) = 0, \quad (40.1)$$

where $D_i^{ab} = \delta^{ab} \partial_i - g_s f^{abc} A_i^c$ is the adjoint covariant derivative, $\hat{E}^{ai}$ is the colour-electric field operator, and $\hat{\rho}^a = \bar{\psi} T^a \psi + f^{abc} A_i^b \Pi^{ci}$ is the colour charge density (quarks + gluons).

The Gauss law generates time-independent gauge transformations:

$$U[\alpha] = \exp\left(i \int d^3\mathbf{x} \alpha^a(\mathbf{x}) \hat{G}^a(\mathbf{x})\right). \quad (40.2)$$

Physical states are gauge-invariant:

$$\hat{G}^a(\mathbf{x}) |\text{phys}\rangle = 0 \quad \text{for all } a, \mathbf{x}. \quad (40.3)$$

This is the Hamiltonian formulation of gauge invariance: the physical Fock space $\mathcal{F}_a^{\text{SM,phys}}$ is the subspace of $\mathcal{F}_a^{\text{SM}}$ satisfying all Gauss law constraints simultaneously.

Physical states are colour singlets. The Gauss law requires the total colour charge $\hat{Q}_{\text{tot}}^a = \int d^3\mathbf{x} \hat{\rho}^a = 0$ in any physical state. Free quarks and free gluons are not in the physical Hilbert space; only colour-singlet combinations (baryons, mesons, glueballs) are. This is the Hamiltonian statement of colour confinement as a selection rule: the Fock space contains all states, but the physical subspace is the colour-singlet sector.

Choosing a physical gauge. In Coulomb gauge ($\nabla \cdot \mathbf{A}^a = 0$), the transverse components of $\mathbf{A}^a$ are the two physical gluon polarisations. The temporal component A_0^a is determined by the Gauss law as a functional of the physical fields:

$$A_0^a(\mathbf{x}) = g_s \int d^3\mathbf{y} \mathcal{G}^{ab}(\mathbf{x}, \mathbf{y}; \mathbf{A}) \rho^b(\mathbf{y}), \quad (40.4)$$

where $\mathcal{G}^{ab} = (-\nabla \cdot D^{ab})^{-1}$. Substituting into the Hamiltonian generates the instantaneous colour-Coulomb interaction:

$$H_{\text{Coulomb}} = \frac{g_s^2}{2} \int \frac{d^3\mathbf{k}}{(2\pi)^3} \rho^a(-\mathbf{k}) \frac{1}{\mathbf{k}^2} [1 + O(g_s^2)] \rho^a(\mathbf{k}). \quad (40.5)$$

The Hamiltonian in Coulomb gauge has no unphysical degrees of freedom and no ghost fields.

40.2 Where the ghost contribution goes

In the path integral (Lorenz gauge), the QCD beta function decomposes as:

$$b_3 = \underbrace{\frac{11N_c}{3}}_{\text{all gluon pols.}} - \underbrace{\frac{N_c}{3}}_{\text{ghost subtraction}} - \underbrace{\frac{4N_f T_F}{3}}_{\text{quark loop}} = 7. \quad (40.6)$$

In the Hamiltonian (Coulomb gauge), the same total rearranges:

Contribution	Path integral	Hamiltonian
All d gluon pols. (overcounting)	$+\frac{11N_c}{3}$	Not present
Ghost subtraction	$-\frac{N_c}{3}$	Not present
Physical transverse gluons (2 pols.)	(implicit: $\frac{11-1}{3}N_c$)	$+\frac{10N_c}{3}$
Coulomb interaction	(inside gluon propagator)	$+\frac{N_c}{3}$
Quark loop	$-\frac{4N_f T_F}{3}$	$-\frac{4N_f T_F}{3}$
Total b_3	7	7

The ghost loop does not add new physics: it subtracts the overcounting from including all d gluon polarisations. In the Hamiltonian approach, only the 2 physical transverse polarisations appear, contributing $10N_c/3$ directly. The remaining $N_c/3$ comes from the instantaneous Coulomb interaction (40.5)—the Hamiltonian counterpart of the ghost loop. The total is gauge-invariant; the decomposition is not.

Ghost fields are path-integral bookkeeping devices. They have no Fock-space interpretation, violate spin-statistics if treated as physical particles, and never produce detector signals. BRST invariance of the path integral is entirely captured by the Gauss law constraint (40.3) in the Hamiltonian.

40.3 The physical Standard Model Fock space

With ghosts removed, the physical SM Fock space at lattice spacing a ($\Lambda = \pi/a$):

$$\mathcal{F}_a^{\text{SM,phys}} = \mathcal{F}_a^{G,\perp} \otimes \mathcal{F}_a^{W,\perp} \otimes \mathcal{F}_a^{B,\perp} \otimes \mathcal{F}_a^\phi \otimes \mathcal{F}_a^q \otimes \mathcal{F}_a^\ell, \quad (40.7)$$

where $\perp$ indicates only the 2 transverse physical polarisations per gauge boson.

Field	Lorentz	$SU(3)_c$	$SU(2)_L$	Stat.	Physical d.o.f.
$G_{\mu,\perp}^a$ (gluon)	$(\frac{1}{2}, \frac{1}{2})$	8	1	B	$8 \times 2 = 16$
$W_{\mu,\perp}^i$ (W boson)	$(\frac{1}{2}, \frac{1}{2})$	1	3	B	$3 \times 2 = 6$
$B_{\mu,\perp}$ (B boson)	$(\frac{1}{2}, \frac{1}{2})$	1	1	B	$1 \times 2 = 2$
ϕ^i (Higgs)	$(0, 0)$	1	2	B	4 real scalars
Quarks q_L, u_R, d_R	$(\frac{1}{2}, 0)$ $(0, \frac{1}{2})$	$\oplus$ 3	2/1	F	$3 \times 3 \times 2 \times 3 = 54$
Leptons ℓ_L, e_R	$(\frac{1}{2}, 0)$ $(0, \frac{1}{2})$	$\oplus$ 1	2/1	F	$3 \times 2 \times 3 = 18$
Total physical d.o.f.					28 bosonic + 72 fermionic

No ghost fields. No unphysical polarisations. Every degree of freedom in $\mathcal{F}_a^{\text{SM,phys}}$ is a physical excitation that can in principle be detected.

The SM engine is a finite lattice. As for the scalar field (Section 20), the Fock-layer engine is a spatial torus $\mathbb{T}^3$ with N_s^3 sites and spacing a , with time evolving continuously (open boundary):

$$h_t = e^{-iH_{\text{SM}}(a)t} h_0, \quad h_t \in \mathcal{F}_a^{\text{SM,phys}}, \quad \|h_t\| = 1. \quad (40.8)$$

The hidden state is one vector of unit norm in a finite-dimensional Fock space (with occupation-number truncation), rotating unitarily under $H_{\text{SM}}(a)$. The Fock-space dimension is finite but inconceivably large: with $N_s = 100$ spatial sites per direction and the SM field content, $\dim \mathcal{F}$ exceeds 10^{10^7} .

The Hamiltonian $H_{\text{SM}}(a)$ is a Hermitian operator on $\mathcal{F}_a^{\text{SM,phys}}$ with engine parameters $\theta_{\text{SM}}(a)$ —the same 19 parameters as on the path-integral side. The lattice dispersion relation replaces $|\mathbf{k}|^2$ with $(2/a)^2 \sum_i \sin^2(k_i a/2)$; all momentum sums are restricted to the Brillouin zone $|k_i| < \pi/a$; every matrix element is finite. The engine runs.

40.4 The Standard Model Hamiltonian

On $\mathcal{F}_a^{\text{SM,phys}}$ in Coulomb gauge:

$$H_{\text{SM}}(a) = H_{\text{free}}(a) + H_{\text{gauge}}(a) + H_{\text{Yukawa}}(a) + H_{\text{Higgs}}(a) + H_{\text{Coulomb}}(a). \quad (40.9)$$

H_{Coulomb} replaces the ghost Hamiltonian: it is the instantaneous colour-Coulomb interaction (40.5).

The free Hamiltonian is diagonal in the Fock basis: $H_{\text{free}} = \sum_{\text{species}} \int_{|\mathbf{k}| < \pi/a} \omega_{\mathbf{k},\text{sp}} \hat{n}_{\mathbf{k},\text{sp}}$ with $\omega_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m_{\text{sp}}^2(a)}$ and the sum over physical modes only.

The interaction vertices are the same as in section 38.2, translated from Lagrangian to Hamiltonian language: each vertex is a polynomial of creation and annihilation operators with momentum-conservation delta functions and coupling constants $g_s(a)$, $g(a)$, $g'(a)$, $y_f(a)$, $\lambda(a)$. The key structural difference from the path integral: the three-gluon and four-gluon vertices involve only transverse polarisation vectors $\epsilon_{\mathbf{k},\lambda}^\mu$ ($\lambda = 1, 2$), and the Coulomb interaction adds an instantaneous four-point interaction between colour charges.

40.5 Feshbach–Löwdin and the SM beta functions

The Feshbach–Löwdin partition (section 20.5) applied to the SM: decompose $\mathcal{F}_a^{\text{SM,phys}}$ into low-momentum P and high-momentum Q sectors, compute H_{eff} on the P sector by integrating

out the shell.

Running of g_s : three channels. The Feshbach correction to the gluon kinetic term receives contributions from three channels in the physical Hamiltonian:

Channel 1: transverse gluon loop (three- and four-gluon vertices with physical polarisations): $\delta(1/g_s^2)_\perp = +(10N_c/3) \cdot \ln(\Lambda/\Lambda')/(16\pi^2)$.

Channel 2: Coulomb interaction ($\rho^a \rho^a / \mathbf{k}^2$ between the shell and the low-momentum sector): $\delta(1/g_s^2)_{\text{Coul}} = +(N_c/3) \cdot \ln(\Lambda/\Lambda')/(16\pi^2)$.

Channel 3: quark loop (screening): $\delta(1/g_s^2)_q = -(4N_f T_F/3) \cdot \ln(\Lambda/\Lambda')/(16\pi^2)$.

Total:

$$b_3 = \frac{10N_c}{3} + \frac{N_c}{3} - \frac{4N_f T_F}{3} = \frac{11N_c}{3} - \frac{4N_f T_F}{3} = 7, \quad (40.10)$$

identical to the path-integral result (39.16).

Running of g and g' . $SU(2)_L$: W - W - W self-interaction (antiscreening) + Coulomb for $SU(2)$ + fermion and Higgs loops (screening): $b_2 = 19/6$, giving $\beta_g = -19g^3/(96\pi^2)$.

$U(1)_Y$: no self-interaction, no Coulomb channel (Abelian), only fermion and scalar screening: $b_1 = -41/6$, giving $\beta_{g'} = +41g'^3/(96\pi^2)$. The absence of the Coulomb channel for $U(1)_Y$ is the Hamiltonian statement that the hypercharge group is Abelian.

40.6 The rendering-stability workflow: Hamiltonian side

The three-step workflow applies on the Hamiltonian side exactly as on the path-integral side, with the Feshbach formula replacing the Feynman-diagram calculation.

Step 1: compute. Fix a and $\theta_{\text{SM}}(a)$. Compute Born-layer observables from $H_{\text{SM}}(a)$: masses from the spectrum of H , scattering amplitudes from the Dyson series $S = T \exp(-i \int H_{\text{int}}^I dt)$, decay widths from the imaginary part of the self-energy. Each observable is a function of $\theta_{\text{SM}}(a)$ and a .

Step 2: impose. The 19 rendering conditions fix the 19 engine parameters. The Feshbach formula gives the explicit running:

Gauge couplings: the Feshbach shell integral reproduces the one-loop beta functions. For $SU(3)_c$: integrating out the momentum shell $\pi/a' < |\mathbf{k}| < \pi/a$ shifts the inverse coupling by:

$$\frac{1}{g_s^2(a')} = \frac{1}{g_s^2(a)} + \frac{b_3}{16\pi^2} \ln \frac{a'}{a}, \quad b_3 = \frac{10N_c}{3} + \frac{N_c}{3} - \frac{4N_f T_F}{3} = 7. \quad (40.11)$$

The $10N_c/3$ comes from transverse gluon loops (physical polarisations only); the $N_c/3$ from the instantaneous Coulomb interaction (the Hamiltonian counterpart of the ghost loop); the $4N_f T_F/3$ from quark screening.

Yukawa couplings: the Feshbach correction to the quark-Higgs vertex gives: $\beta_{y_t} = y_t/(16\pi^2) \cdot (\frac{9}{2}y_t^2 - 8g_s^2 - \frac{9}{4}g^2 - \frac{17}{12}g'^2) + \dots$

Higgs quartic: the Feshbach correction gives: $\beta_\lambda = \frac{1}{16\pi^2}(24\lambda^2 + 12\lambda y_t^2 - 6y_t^4 - 3\lambda(3g^2 + g'^2) + \dots)$.

Higgs mass parameter: the Feshbach self-energy gives $m^2(a) = m_{\text{phys}}^2 - c \cdot (\text{SM couplings})/a^2$, where c involves $6\lambda + 6y_t^2 - \frac{9}{4}g^2 - \frac{3}{4}g'^2$ from $\mathbf{A}_1$.

Step 3: verify. Once 19 conditions are imposed, all higher-point functions are automatically a -independent. The Gauss law constraint on the Hamiltonian side plays the role of BRST on

the path-integral side: it ensures that only gauge-invariant operators appear as counterterms, reducing the independent structures to 19.

Agreement with the path-integral side. Every Born-layer prediction is identical:

Quantity	Path integral	Hamiltonian
β_{g_s}	$\frac{11N_c}{3} - \frac{N_c}{3} - \frac{4N_f T_F}{3}$	$\frac{10N_c}{3} + \frac{N_c}{3} - \frac{4N_f T_F}{3}$
b_3 total	7	7
Ghost sector	Ghost loop ($-N_c/3$)	Coulomb int. ($+N_c/3$)
$m^2(a)$	$\mathbf{A}_1$ coefficient	Feshbach self-energy
$\theta_{\text{SM}}(a)$	Same 19 numbers	Same 19 numbers
Born outputs	Same	Same

The decomposition into channels is a Fock-layer bookkeeping choice (ghosts vs. Coulomb); the Born-layer totals are physical and agree. No experiment can distinguish the ghost contribution from the Coulomb contribution—they are two descriptions of the same physics.

40.7 Hamiltonian versus path integral: the full comparison

	Hamiltonian	Path integral
<i>Gauge invariance</i>	Gauss law constraint on $\mathcal{F}^{\text{phys}}$	Faddeev–Popov + BRST on $\int \mathcal{D}\Phi$
<i>Ghost fields</i>	Absent	Required for unitarity of S -matrix
<i>Unphysical pols.</i>	Absent (Coulomb gauge)	Present; subtracted by ghosts
<i>Gluon d.o.f.</i>	2 transverse + Coulomb interaction	d covariant – ghost
<i>Fock space</i>	$\mathcal{F}_a^{\text{SM,phys}}$: physical only	$\mathcal{F}_a^{\text{SM,ext}}$: includes ghosts
<i>Beta function</i>	Feshbach: $\frac{10N_c}{3} + \frac{N_c}{3} - \frac{4N_f T_F}{3}$	Heat kernel: $\frac{11N_c}{3} - \frac{N_c}{3} - \frac{4N_f T_F}{3}$
<i>Total b_3</i>	7	7
<i>Lorentz covariance</i>	Not manifest (Coulomb gauge breaks it)	Manifest (covariant gauge)
<i>Renormalization</i>	Feshbach H_{eff} , Wegner flow	$\log Z$, effective action Γ
<i>Stability condition</i>	$\mathcal{O}_{\text{Born}} = \mathcal{O}_{\text{phys}} + o(a/\lambda)$	Same
<i>Non-perturbative</i>	DMRG, tensor networks, quantum sim. (no sign problem)	QMC (sign problem for fermions at finite μ)
<i>Practical strength</i>	Conceptual clarity; no ghosts; direct Fock-space access	Computational efficiency; gauge-invariant regularisation; manifest Lorentz

The decomposition differs; the total agrees. The beta function coefficient $b_3 = 7$ is a Born-layer observable: it determines the measurable running of α_s with energy. Its decomposition into channels (transverse gluon + Coulomb + quark in the Hamiltonian; all-polarisation gluon – ghost + quark in the path integral) is a Fock-layer bookkeeping choice, not a physical prediction. No experiment can measure the “ghost contribution” to the beta function, because no experiment can distinguish the ghost loop from the Coulomb interaction—they are two descriptions of the same physics (the effect of the eliminated unphysical polarisations).

When to use which. The path integral excels at perturbative calculations in gauge theories: Lorentz covariance is manifest, dimensional regularization preserves gauge invariance automatically, and the Feynman-diagram technology is highly developed. Nearly all precision SM predictions ($(g-2)_e$, Higgs production cross sections, electroweak precision observables) are computed in the path-integral framework.

The Hamiltonian excels at conceptual clarity and non-perturbative problems. It makes the physical degrees of freedom explicit (no ghosts, no unphysical polarisations), connects directly to the Fock space and the Born rule, and provides access to non-perturbative methods (exact diagonalisation, DMRG, tensor networks, quantum simulation) that do not have a sign problem. For lattice QCD at finite baryon density, where the Euclidean path integral suffers from the sign problem, Hamiltonian methods are the most promising avenue.

Both are Fock-layer tools. The Hamiltonian $H_{\text{SM}}(a)$ and the path integral $\int \mathcal{D}\Phi e^{-S_{\text{SM}}}$ are two computational strategies for the same Fock-layer engine. The Born layer does not care which strategy the engine uses: it renders the same 19 physical quantities, the same cross sections, the same masses, the same mixing angles. The choice between Hamiltonian and path integral is a choice of computational tool, not of physics.

The Hamiltonian is conceptually primary (it defines the engine), but the path integral is computationally indispensable (it provides the most efficient tools for perturbative gauge-theory calculations). A complete understanding of the Standard Model requires fluency in both.

40.8 The Standard Model in the Fock-Born framework: a summary

Layer	What it contains	What it does NOT contain
Fock	$H_{\text{SM}}(a)$ on $\mathcal{F}_a^{\text{SM,phys}}$; engine parameters $\theta_{\text{SM}}(a)$; correlators $G^{(n)}$; vertex functions $\Gamma^{(n)}$; S -matrix elements $\langle f S i\rangle$	Events; detector clicks; “emission”; “exchange”; particles travelling; ghosts as physical entities
Born	Cross sections $\sigma = \mathcal{M} ^2$; masses m_{phys} (from spectrometers); decay widths Γ (from Breit–Wigner fits); mixing angles (from amplitude ratios); asymmetries (from event-rate differences)	Propagators; vertex functions; self-energies; engine parameters $\theta(a)$; ghost loops; Feynman diagrams

The Standard Model is one engine $H_{\text{SM}}(a)$ with 19 engine parameters $\theta_{\text{SM}}(a)$, operating on one Fock space $\mathcal{F}_a^{\text{SM,phys}}$, producing one hidden state h_t that rotates unitarily. All Born-layer predictions—every cross section, mass, decay width, mixing angle, and CP asymmetry ever measured—follow from this single engine via the rendering-stability condition and the Born rule. The engine parameters depend on the lattice spacing a ; the Born outputs do not. This is the complete content of SM renormalization.

41 Cataloging the Standard Model

41.1 A field guide to the particles: terminology and classification

The Standard Model has a small number of fundamental fields but an enormous number of bound states built from them. The terminology has accumulated over seven decades of discovery, with names drawn from Greek, Latin, Japanese, Hebrew, and the whims of individual physicists. This subsection is a reference guide.

The fundamental fields

The Standard Model contains 17 fundamental fields (not counting antiparticles separately), organised by spin:

Spin	Name	Symbol	Charge	Colour	Mass
<i>Spin-$\frac{1}{2}$ fermions: quarks (feel the strong force)</i>					
$\frac{1}{2}$	Up	u	$+\frac{2}{3}$	3	2.2 MeV
$\frac{1}{2}$	Down	d	$-\frac{1}{3}$	3	4.7 MeV
$\frac{1}{2}$	Charm	c	$+\frac{2}{3}$	3	1.27 GeV
$\frac{1}{2}$	Strange	s	$-\frac{1}{3}$	3	93 MeV
$\frac{1}{2}$	Top	t	$+\frac{2}{3}$	3	173 GeV
$\frac{1}{2}$	Bottom	b	$-\frac{1}{3}$	3	4.18 GeV
<i>Spin-$\frac{1}{2}$ fermions: leptons (do not feel the strong force)</i>					
$\frac{1}{2}$	Electron	e^-	-1	1	0.511 MeV
$\frac{1}{2}$	Electron neutrino	ν_e	0	1	< 1 eV
$\frac{1}{2}$	Muon	μ^-	-1	1	105.7 MeV
$\frac{1}{2}$	Muon neutrino	ν_μ	0	1	< 1 eV
$\frac{1}{2}$	Tau	τ^-	-1	1	1.777 GeV
$\frac{1}{2}$	Tau neutrino	ν_τ	0	1	< 1 eV
<i>Spin-1 bosons: gauge fields (force carriers)</i>					
1	Photon	γ	0	1	0
1	W bosons	$W^\pm$	± 1	1	80.4 GeV
1	Z boson	Z^0	0	1	91.2 GeV
1	Gluons (8)	g	0	8	0
<i>Spin-0 boson: scalar</i>					
0	Higgs	H	0	1	125.1 GeV

Every fundamental fermion has an antiparticle with the same mass and opposite quantum numbers (charge, colour, baryon/lepton number): $\bar{u}$, $\bar{d}$, e^+ , $\bar{\nu}_e$, etc. The photon, Z , gluons, and Higgs are their own antiparticles.

Generations. The quarks and leptons come in three copies (*generations* or *families*), identical in gauge quantum numbers but differing in mass:

	Generation I	Generation II	Generation III
Quarks	u, d	c, s	t, b
Leptons	e^-, ν_e	μ^-, ν_μ	τ^-, ν_τ

All ordinary matter is made from Generation I (u, d, e^-, ν_e). Generations II and III are heavier copies that decay rapidly to Generation I. Why there are exactly three generations is unexplained.

The classification of bound states: hadrons

The bound states of quarks and gluons are called *hadrons* (from the Greek *hadros*, “thick” or “strong”). All hadrons are colour singlets: the strong force confines quarks into colourless combinations.

Hadrons are classified by their quark content:

Type	Quark content	Baryon no.	Statistics	Colour structure
Meson	$q\bar{q}$	0	Boson	$\delta^i_j/\sqrt{3}$ (colour-anticolour)
Baryon	qqq	+1	Fermion	$\epsilon^{ijk}/\sqrt{6}$ (antisymmetric)
Antibaryon	$\bar{q}\bar{q}\bar{q}$	-1	Fermion	$\epsilon_{ijk}/\sqrt{6}$
Tetraquark	$qq\bar{q}\bar{q}$	0	Boson	Various colour singlets
Pentaquark	$qqqq\bar{q}$	+1	Fermion	Various colour singlets
Glueball	gg, ggg	0	Boson	Pure glue, colour singlet

Tetraquarks, pentaquarks, and glueballs are *exotic hadrons*: they have been observed (tetraquarks and pentaquarks at LHCb, beginning 2015) but are much rarer than mesons and baryons. Glueballs are predicted by lattice QCD but not yet unambiguously identified experimentally.

Hadrons are further classified by spin and parity

Each hadron has a definite spin J and parity P (and, for flavour-neutral hadrons, charge conjugation C), written as J^{PC} or J^P . The spin-parity is determined by the orbital angular momentum L and the intrinsic spins of the quarks.

The meson zoo

Mesons ($q\bar{q}$) are classified by their J^{PC} quantum numbers and their flavour content.

Pseudoscalar mesons ($J^{PC} = 0^{-+}$, $L = 0$, spins antiparallel):

Name	Symbol	Content	Mass (MeV)
Pion	π^+, π^0, π^-	$u\bar{d}, (u\bar{u} - d\bar{d})/\sqrt{2}, d\bar{u}$	140/135/140
Kaon	K^+, K^0	$u\bar{s}, d\bar{s}$	494/498
Antikaon	$K^-, \bar{K}^0$	$s\bar{u}, s\bar{d}$	494/498
Eta	η	$(u\bar{u} + d\bar{d} - 2s\bar{s})/\sqrt{6}$	548
Eta prime	η'	$(u\bar{u} + d\bar{d} + s\bar{s})/\sqrt{3}$	958
D meson	D^+, D^0	$c\bar{d}, c\bar{u}$	1870/1865
D _s meson	D_s^+	$c\bar{s}$	1968
B meson	B^+, B^0	$u\bar{b}, d\bar{b}$	5279/5280
B _s meson	B_s^0	$s\bar{b}$	5367
B _c meson	B_c^+	$c\bar{b}$	6275

The pions are the lightest hadrons and the pseudo-Goldstone bosons of chiral symmetry breaking. The kaons carry strangeness ($S = \pm 1$). The D mesons carry charm ($C = \pm 1$); the B mesons carry bottom ($B' = \pm 1$, where B' is the bottom quantum number, distinct from baryon number B).

Vector mesons ($J^{PC} = 1^{--}$, $L = 0$, spins parallel):

Name	Symbol	Content	Mass (MeV)
Rho	ρ^+, ρ^0, ρ^-	$u\bar{d}, (u\bar{u} - d\bar{d})/\sqrt{2}, d\bar{u}$	775
Omega	ω	$(u\bar{u} + d\bar{d})/\sqrt{2}$	782
Phi	ϕ	$s\bar{s}$	1020
K^*	K^{*+}, K^{*0}	$u\bar{s}, d\bar{s}$	892
J/ψ	J/ψ	$c\bar{c}$	3097
Υ	Υ	$b\bar{b}$	9460
D^*	D^{*+}, D^{*0}	$c\bar{d}, c\bar{u}$	2010/2007
B^*	B^{*+}, B^{*0}	$u\bar{b}, d\bar{b}$	5325

The ρ is the spin-1 partner of the pion. The J/ψ ($c\bar{c}$) and Υ ($b\bar{b}$) are the *quarkonia*: heavy quark-antiquark bound states analogous to positronium.

Scalar mesons ($J^{PC} = 0^{++}$): $f_0(500)$ (the σ , broad), $f_0(980)$, $a_0(980)$, $K_0^*(700)$ (the κ). These are controversial: some may be $q\bar{q}$, others $qq\bar{q}\bar{q}$ tetraquarks or $K\bar{K}$ molecules.

Tensor mesons ($J^{PC} = 2^{++}$): $f_2(1270)$, $a_2(1320)$, $K_2^*(1430)$.

Charmonium and bottomonium families. The $c\bar{c}$ (*charmonium*) and $b\bar{b}$ (*bottomonium*) systems have a rich spectroscopy of excited states, labelled by J^{PC} :

J^{PC}	Spectroscopic	Charmonium	Bottomonium	Name
0^{-+}	$^1S_0$	$\eta_c(2984)$	$\eta_b(9399)$	Pseudoscalar
1^{--}	$^3S_1$	$J/\psi(3097)$	$\Upsilon(9460)$	Vector
0^{++}	$^3P_0$	$\chi_{c0}(3415)$	$\chi_{b0}(9859)$	Scalar
1^{++}	$^3P_1$	$\chi_{c1}(3511)$	$\chi_{b1}(9893)$	Axial vector
2^{++}	$^3P_2$	$\chi_{c2}(3556)$	$\chi_{b2}(9912)$	Tensor
1^{+-}	$^1P_1$	$h_c(3526)$	$h_b(9899)$	Axial vector

The spectroscopic notation $^{2S+1}L_J$ labels the spin S , orbital angular momentum L ($S, P, D, \dots$

for $L = 0, 1, 2, \dots$), and total angular momentum J .

The baryon zoo

Baryons (qqq) are classified by their quark flavour content, spin, and parity.

The spin- $\frac{1}{2}$ baryon octet ($J^P = \frac{1}{2}^+, L = 0$):

Name	Symbol	Content	I	S	Mass (MeV)
Proton	p	uud	$\frac{1}{2}$	0	938
Neutron	n	udd	$\frac{1}{2}$	0	940
Lambda	Λ	uds	0	-1	1116
Sigma	$\Sigma^+, \Sigma^0, \Sigma^-$	uus, uds, dds	1	-1	1189/1192/1197
Xi (cascade)	Ξ^0, Ξ^-	uss, dss	$\frac{1}{2}$	-2	1315/1322

The Λ and Σ^0 have the same quark content (uds) but differ in isospin: Λ has $I = 0$, Σ^0 has $I = 1$. They are different eigenstates of the $SU(3)$ flavour Hamiltonian.

The spin- $\frac{3}{2}$ baryon decuplet ($J^P = \frac{3}{2}^+, L = 0$):

Name	Symbol	Content	I	S	Mass (MeV)
Delta	$\Delta^{++}, \Delta^+, \Delta^0, \Delta^-$	uuu, uud, udd, ddd	$\frac{3}{2}$	0	1232
Sigma star	$\Sigma^{*+}, \Sigma^{*0}, \Sigma^{*-}$	uus, uds, dds	1	-1	1385
Xi star	Ξ^{*0}, Ξ^{*-}	uss, dss	$\frac{1}{2}$	-2	1533
Omega	Ω^-	sss	0	-3	1672

The $\Omega^- = sss$ was the particle predicted by Gell-Mann from the Eightfold Way and discovered at Brookhaven in 1964.

Charmed and bottom baryons. Replacing one or more light quarks (u, d, s) with a heavy quark (c, b) gives:

Name	Symbol	Content	Mass (MeV)	Charm/Bottom
Lambda _c	Λ_c^+	udc	2286	$C = +1$
Sigma _c	$\Sigma_c^{++,+,0}$	uuc, udc, ddc	2455	$C = +1$
Xi _c	$\Xi_c^{+,0}$	usc, dsc	2470	$C = +1$
Omega _c	Ω_c^0	ssc	2695	$C = +1$
Xi _{cc}	Ξ_{cc}^{++}	ucc	3621	$C = +2$
Lambda _b	Λ_b^0	udb	5620	$B' = -1$
Xi _b	$\Xi_b^{0,-}$	usb, dsb	5792	$B' = -1$
Omega _b	Ω_b^-	ssb	6046	$B' = -1$

The naming conventions

The names follow patterns, though with historical irregularities:

Mesons are named by their J^{PC} and flavour:

- Light unflavoured (u, d only): π (0^{-+}), ρ (1^{--}), ω (1^{--} , isospin 0), η (0^{-+} , isospin 0), f (even J^{++}), a (even J^{++} , isospin 1), b (1^{+-}), h (1^{+-} , isospin 0).

- Strange ($s\bar{q}$ or $q\bar{s}$): K (0^-), K^* (1^-), K_0^* (0^+), K_2^* (2^+).
- Charmed ($c\bar{q}$): D (0^-), D^* (1^-), D_s ($c\bar{s}$), D_s^* .
- Bottom ($b\bar{q}$): B (0^-), B^* (1^-), B_s ($b\bar{s}$), B_c ($b\bar{c}$).
- Charmonium ($c\bar{c}$): η_c , J/ψ , $\psi(2S)$, χ_{cJ} , h_c .
- Bottomonium ($b\bar{b}$): η_b , Υ , χ_{bJ} , h_b .

Baryons are named by strangeness content and isospin:

- $S = 0$: N (nucleon: p , n ; $I = \frac{1}{2}$), Δ ($I = \frac{3}{2}$).
- $S = -1$: Λ ($I = 0$), Σ ($I = 1$).
- $S = -2$: Ξ (cascade; $I = \frac{1}{2}$).
- $S = -3$: Ω ($I = 0$).
- With charm: subscript c (Λ_c , Σ_c , Ξ_c , Ω_c).
- With bottom: subscript b (Λ_b , Ξ_b , Ω_b).
- Excited states: a star (Σ^* , Ξ^*) or a mass in parentheses ($N(1440)$, $\Delta(1600)$).

Exotic hadrons

States beyond the conventional $q\bar{q}$ and qqq classifications:

Type	Content	Examples	Mass (MeV)	Discovered
Tetraquark	$qq\bar{q}\bar{q}$	$X(3872)$, $Z_c(3900)^+$	3872, 3900	Belle 2003, BES III 2013
Pentaquark	$qqqq\bar{q}$	$P_c(4312)^+$, $P_c(4450)^+$	4312, 4450	LHCb 2015, 2019
Glueball	gg , ggg	Candidates: $f_0(1710)?$	~ 1700	Not confirmed
Hybrid	$q\bar{q}g$	Candidates: $\pi_1(1600)?$	~ 1600	Under study

The $X(3872)$, discovered by the Belle collaboration in 2003, was the first established exotic hadron. It sits at the $D^0\bar{D}^{*0}$ threshold and is debated: a $c\bar{c}u\bar{u}$ tetraquark, a $D\bar{D}^*$ molecule, or a mixture. LHCb has since discovered a zoo of exotic states (X , Y , Z , P_c , T_{cc}), many of which sit near meson-meson or meson-baryon thresholds, suggesting molecular interpretations.

Collective terminology

Term	Meaning
Hadron	Any bound state of quarks and/or gluons (meson, baryon, exotic)
Lepton	Fundamental fermion that does not feel the strong force ($e, \mu, \tau, \nu_e, \nu_\mu, \nu_\tau$)
Fermion	Any particle with half-integer spin (quarks, leptons, baryons)
Boson	Any particle with integer spin (photon, W, Z , gluon, Higgs, mesons)
Baryon	Hadron with baryon number $B = +1$ (three quarks)
Meson	Hadron with baryon number $B = 0$ (quark-antiquark)
Nucleon	Proton or neutron (N)
Hyperon	Baryon with strangeness $S \neq 0$ ($\Lambda, \Sigma, \Xi, \Omega$)
Quarkonium	$q\bar{q}$ bound state of a heavy quark and its own antiquark: charmonium ($c\bar{c}$) or bottomonium ($b\bar{b}$)
Resonance	Short-lived hadron that decays strongly (typical lifetime $\sim 10^{-23}$ s); appears as a peak in a scattering cross section
Stable hadron	Hadron that does not decay strongly: p, n (decays weakly, $\tau \sim 880$ s), $\pi^\pm$ (weakly, $\tau \sim 26$ ns), $K^\pm$ (weakly), Λ (weakly)
Gauge boson	Spin-1 particle mediating a force ($\gamma, W^\pm, Z^0, g$)
Gluon	Gauge boson of $SU(3)_c$ (8 colour states, massless, self-interacting)
Photon	Gauge boson of $U(1)_{\text{em}}$ (massless, no self-interaction)
W/Z bosons	Gauge bosons of $SU(2)_L \times U(1)_Y$ (massive, from Higgs mechanism)
Higgs boson	Spin-0 boson from EWSB; the σ of the electroweak linear sigma model
Graviton	Hypothetical spin-2 boson mediating gravity (not in the SM)
Flavour	The species of quark (u, d, s, c, b, t) or lepton (e, μ, τ)
Colour	The $SU(3)_c$ charge of quarks (red, green, blue) and gluons (eight colour states)
Generation	A complete set of quarks and leptons with the same gauge quantum numbers (I, II, III)
Chirality	The eigenvalue of γ_5 : left-handed (L) or right-handed (R)
Helicity	The projection of spin onto momentum: parallel (R) or antiparallel (L); equals chirality for massless particles
Valence quarks	The quarks that determine the quantum numbers of a hadron (uud for the proton)
Sea quarks	Virtual $q\bar{q}$ pairs in the hadron's Fock decomposition
Parton	Any constituent of a hadron as seen in deep inelastic scattering: valence quarks, sea quarks, and gluons
Jet	A collimated spray of hadrons produced by the fragmentation of a high-energy quark or gluon

The flavour quantum numbers

Each quark flavour carries a conserved (or approximately conserved) quantum number:

Quantum no.	Symbol	Quark	Value for quark	Conserved by
Baryon number	B	All quarks	$+\frac{1}{3}$	Strong, EM, weak
Strangeness	S	s	-1	Strong, EM (not weak)
Charm	C	c	$+1$	Strong, EM (not weak)
Bottomness	B'	b	-1	Strong, EM (not weak)
Topness	T'	t	$+1$	Strong, EM (not weak)
Isospin (I_3)	I_3	u/d	$+\frac{1}{2}/-\frac{1}{2}$	Strong (not EM, weak)

The weak interaction violates all flavour quantum numbers except baryon number and lepton number (which are accidental symmetries of the renormalizable SM Lagrangian). The CKM matrix parametrises the weak transitions between quark flavours.

Why there are so many names

The proliferation of particle names reflects the history of discovery. In the 1940s–1960s, each new particle detected in cosmic rays or accelerators was given an individual name because its relationship to other particles was unknown. The pion, the kaon, the lambda, the sigma, the xi, the omega were named as independent objects—each a surprise, each a mystery.

The Eightfold Way revealed that these names label components of $SU(3)$ multiplets: the “separate” particles are different states of the same representation. The quark model revealed that the multiplets are built from three fundamental objects. QCD revealed that the quarks are bound by gluons in the Fock space.

In the Fock-Born framework: the fundamental fields ($u, d, s, c, b, t, e, \mu, \tau, \nu_e, \nu_\mu, \nu_\tau, \gamma, W^\pm, Z^0, g, H$) are Fock-layer creation operators—the basis vectors of the game engine. The hadrons ($\pi, K, p, n, \Lambda, \Sigma, \Xi, \Omega, J/\psi, \Upsilon, \dots$) are Born-layer collective modes—bound states in the QCD Fock space, each a specific superposition of Fock basis vectors that rotates at a single frequency $mc^2/\hbar$ under the Schrödinger equation. The hundreds of named hadrons are the hundreds of stable patterns in the amplitude distribution $c(\dots; t)$ of the QCD Fock hidden state. The names are labels for the patterns; the game engine knows only quarks and gluons.

41.2 Hadrons as patterns in the QCD Fock space: a formal treatment

The preceding subsection listed the hundreds of named hadrons and stated that they are Born-layer collective modes of the QCD Fock space. We now make this statement precise.

The QCD Fock space

The Fock space of QCD is built from creation operators for quarks and gluons:

$$\begin{aligned}
b_{f,\alpha,\mathbf{k},s}^\dagger &: \text{creates a quark of flavour } f \in \{u, d, s, c, b, t\}, \text{ colour } \alpha \in \{1, 2, 3\}, \text{ momentum } \mathbf{k}, \text{ spin } s, \\
d_{f,\alpha,\mathbf{k},s}^\dagger &: \text{creates an antiquark,} \\
a_{a,\mathbf{k},\lambda}^\dagger &: \text{creates a gluon of colour } a \in \{1, \dots, 8\}, \text{ momentum } \mathbf{k}, \text{ helicity } \lambda \in \{\pm 1\}.
\end{aligned} \tag{41.1}$$

The Fock basis vectors are products of these creation operators acting on the vacuum $|\Omega\rangle$ (the interacting vacuum of QCD, not the free vacuum):

$$|\{n_{f\alpha\mathbf{k}s}, \bar{n}_{f\alpha\mathbf{k}s}, n_{a\mathbf{k}\lambda}\}\rangle = \prod (b^\dagger)^n (d^\dagger)^{\bar{n}} (a^\dagger)^n |\Omega\rangle. \tag{41.2}$$

Each basis vector specifies, for every mode, how many quarks, antiquarks, and gluons are present. The Fock space is the direct sum over all particle numbers:

$$\mathcal{F}_{\text{QCD}} = \bigoplus_{n_q, n_{\bar{q}}, n_g=0}^{\infty} \mathcal{H}^{(n_q, n_{\bar{q}}, n_g)}, \quad (41.3)$$

where $\mathcal{H}^{(n_q, n_{\bar{q}}, n_g)}$ is the Hilbert space of n_q quarks, $n_{\bar{q}}$ antiquarks, and n_g gluons (appropriately symmetrised: antisymmetric in quarks, symmetric in gluons).

The dimension of this Fock space is infinite (continuous momenta, unbounded particle number). For a finitized version (lattice with N^3 spatial sites), the dimension is finite but astronomical.

The QCD Hamiltonian

The QCD Hamiltonian, expressed in terms of the creation and annihilation operators, has the structure:

$$H_{\text{QCD}} = H_0 + H_{\text{int}}, \quad (41.4)$$

where the free Hamiltonian:

$$H_0 = \sum_{f, \alpha, \mathbf{k}, s} \omega_{f\mathbf{k}} b_{f\alpha\mathbf{k}s}^\dagger b_{f\alpha\mathbf{k}s} + \sum_{f, \alpha, \mathbf{k}, s} \omega_{f\mathbf{k}} d_{f\alpha\mathbf{k}s}^\dagger d_{f\alpha\mathbf{k}s} + \sum_{a, \mathbf{k}, \lambda} |\mathbf{k}| a_{a\mathbf{k}\lambda}^\dagger a_{a\mathbf{k}\lambda} \quad (41.5)$$

is diagonal in the Fock basis, with $\omega_{f\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + m_f^2}$. The interaction Hamiltonian contains:

$$\begin{aligned} H_{\text{int}} &= g_s \sum \bar{\psi} \gamma^\mu T^a \psi A_\mu^a && \text{(quark-gluon vertex: } q \rightarrow qg) \\ &+ g_s \sum f^{abc} A_\mu^a A_\nu^b \partial^\mu A^{c\nu} && \text{(three-gluon vertex: } g \rightarrow gg) \\ &+ g_s^2 \sum f^{abe} f^{cde} A^a A^b A^c A^d && \text{(four-gluon vertex: } gg \rightarrow gg). \end{aligned} \quad (41.6)$$

Each vertex changes the Fock basis vector: the quark-gluon vertex moves amplitude from the sector (n_q, n_g) to $(n_q, n_g \pm 1)$; the three-gluon vertex moves between (n_g) and $(n_g \pm 1)$; the four-gluon vertex rearranges gluons.

The key point: H_0 is diagonal in the Fock basis (each basis vector is an eigenstate of H_0). The full Hamiltonian H_{QCD} is *not* diagonal: the interaction H_{int} has off-diagonal matrix elements that connect different Fock basis vectors. The eigenstates of H_{QCD} are therefore *superpositions* of Fock basis vectors—not individual basis vectors.

Hadrons as eigenstates of H_{QCD}

A hadron h with mass m_h is an eigenstate of the full QCD Hamiltonian:

$$H_{\text{QCD}} |h, \mathbf{P}\rangle = E_h(\mathbf{P}) |h, \mathbf{P}\rangle, \quad E_h(\mathbf{P}) = \sqrt{|\mathbf{P}|^2 + m_h^2}, \quad (41.7)$$

where $\mathbf{P}$ is the total three-momentum. In the Fock basis, this eigenstate is a superposition:

$$|h, \mathbf{P}\rangle = \sum_{\{f_i, \alpha_i, \mathbf{k}_i, s_i\}} \phi_h^{(n_q, n_{\bar{q}}, n_g)}(\{f_i, \alpha_i, \mathbf{k}_i, s_i\}) \prod_i b_{f_i \alpha_i \mathbf{k}_i s_i}^\dagger \prod_j d_{f_j \alpha_j \mathbf{k}_j s_j}^\dagger \prod_k a_{a_k \mathbf{k}_k \lambda_k}^\dagger |\Omega\rangle \quad (41.8)$$

summed over all Fock sectors $(n_q, n_{\bar{q}}, n_g)$, subject to the constraints:

$$\text{total momentum: } \sum_i \mathbf{k}_i = \mathbf{P}, \quad (41.9)$$

$$\text{colour singlet: } \text{total colour} = \mathbf{1} \text{ (trivial rep of } SU(3)_c), \quad (41.10)$$

$$\text{quantum numbers: } \text{total } J^{PC}, I, S, C, B' = \text{those of } h. \quad (41.11)$$

The wavefunction $\phi_h^{(n_q, n_{\bar{q}}, n_g)}$ is the amplitude for the hadron to be found in a specific Fock configuration. It depends on all the momenta, spins, flavours, and colours of the constituents, and it is determined by the eigenvalue equation (41.7)—which is, in general, a non-perturbative problem solvable only by lattice QCD.

The Fock decomposition is infinite. The sum in (41.8) runs over *all* Fock sectors: $n_q = 3, n_g = 0$ (valence quarks alone, for a baryon); $n_q = 3, n_g = 1$ (valence quarks plus one gluon); $n_q = 5, n_{\bar{q}} = 2, n_g = 0$ (valence quarks plus a sea $q\bar{q}$ pair); and so on, to arbitrarily high particle number. The hadron is not a finite-body system; it is a superposition over an infinite tower of Fock sectors.

For a light hadron (pion, proton): the higher Fock sectors carry substantial weight. More than 99% of the proton’s mass comes from the gluon field energy and the kinetic energy of the quarks, not from the quark masses. The “valence” sector ($uud, n_g = 0$) contributes only a few percent of the proton’s momentum: the remainder is carried by gluons and sea quarks.

For a heavy quarkonium ($J/\psi, \Upsilon$): the lowest Fock sector ($c\bar{c}$ or $b\bar{b}, n_g = 0$) dominates, with corrections of order $\alpha_s(m_Q) \approx 0.2\text{--}0.3$.

Each hadron rotates at a single frequency

The Schrödinger equation $\dot{h}_t = -iH_{\text{QCD}}h_t$ evolves the Fock hidden state unitarily. Because the hadron $|h, \mathbf{P}\rangle$ is an eigenstate of H_{QCD} with eigenvalue E_h , it evolves as:

$$|h, \mathbf{P}; t\rangle = e^{-iE_h t} |h, \mathbf{P}; 0\rangle. \quad (41.12)$$

The *entire superposition*—all the Fock basis vectors with their amplitudes—rotates rigidly in the Fock space at a single angular frequency:

$$\omega_h = \frac{E_h}{\hbar} = \frac{\sqrt{|\mathbf{P}|^2 c^2 + m_h^2 c^4}}{\hbar}. \quad (41.13)$$

In the rest frame ($\mathbf{P} = \mathbf{0}$): $\omega_h = m_h c^2 / \hbar$. The proton rotates at $\omega_p = m_p c^2 / \hbar \approx 1.4 \times 10^{24} \text{ s}^{-1}$; the pion at $\omega_\pi \approx 2.1 \times 10^{23} \text{ s}^{-1}$; the J/ψ at $\omega_{J/\psi} \approx 4.7 \times 10^{24} \text{ s}^{-1}$.

The individual Fock basis vectors in the superposition are *not* eigenstates of H_{QCD} : they rotate at different frequencies under H_0 (each with $\omega = E_{H_0} / \hbar$), and the interaction H_{int} continuously reshuffles amplitude among them. But the specific superposition that is the hadron is an eigenstate: the reshuffling is exactly balanced, and the pattern as a whole rotates rigidly.

This is what it means for a hadron to be a “stable pattern in the amplitude distribution”: the amplitudes $\phi_h(\{f_i, \alpha_i, \mathbf{k}_i, s_i\})$ are individually changing under H_0 (each Fock basis vector acquires a different phase), but H_{int} compensates exactly, so that the overall superposition acquires a single phase $e^{-iE_h t}$. The pattern is locked in by the interplay of H_0 and H_{int} .

The Born observable that defines each hadron

Each hadron is defined by a Born observable: the invariant mass in a specific quantum-number channel. For a general hadron h with quantum numbers $\mathcal{Q} = (J^{PC}, I, I_3, S, C, B', B)$:

$$\hat{\mathcal{O}}_h = \hat{M}^2|_{\mathcal{Q}, \text{ colour singlet}}. \quad (41.14)$$

The eigenvalue m_h^2 labels the hadron. The degeneracy of this eigenvalue in $\mathcal{F}_{\text{QCD}}$ is the number of Fock configurations compatible with quantum numbers $\mathcal{Q}$, total invariant mass m_h , and colour singlet—an enormous number, encompassing all the internal momentum distributions, all the gluon and sea-quark configurations, all the Fock sectors.

The eigensubspace factorises:

$$\mathcal{H}_{m_h^2, \mathcal{Q}} \approx |h, \mathbf{P}\rangle_{\text{hadron}} \otimes \mathcal{H}_{\text{internal}}^{(h)}, \quad (41.15)$$

where $|h, \mathbf{P}\rangle$ is the hadron state (labelled by centre-of-mass momentum $\mathbf{P}$) and $\mathcal{H}_{\text{internal}}^{(h)}$ is the space of internal configurations: all the ways to distribute momentum, spin, colour, and flavour among the quarks and gluons inside the hadron. The Born probability for finding hadron h with momentum $\mathbf{P}$:

$$P(h, \mathbf{P}) = \|P_{m_h^2, \mathcal{Q}, \mathbf{P}} h_t\|^2, \quad (41.16)$$

is the squared length of the projected superposition, with the OR sum running over all internal configurations.

The spectrum as a catalogue of eigenfrequencies

The hadron spectrum—the list of all hadrons with their masses, quantum numbers, and widths—is the catalogue of eigenfrequencies of H_{QCD} in the colour-singlet sector.

Stable hadrons are eigenstates of the full Hamiltonian $H = H_{\text{QCD}} + H_{\text{weak}} + H_{\text{EM}}$ with zero width in the strong interaction: they can only decay weakly or electromagnetically. The proton is stable (lifetime $> 10^{34}$ years); the neutron decays weakly ($\tau \approx 880$ s); the pion decays weakly ($\pi^\pm$, $\tau \approx 26$ ns) or electromagnetically ($\pi^0 \rightarrow \gamma\gamma$, $\tau \approx 8 \times 10^{-17}$ s). In the QCD Fock space, these are exact eigenstates of H_{QCD} : the pattern rotates at a single frequency forever (in the absence of weak and EM interactions).

Resonances are eigenstates of H_{QCD} with nonzero width: they can decay strongly to lighter hadrons. The ρ meson ($\Gamma = 149$ MeV, $\tau \approx 4 \times 10^{-24}$ s), the Δ baryon ($\Gamma = 117$ MeV), and the hundreds of excited hadrons listed by the Particle Data Group are resonances.

A resonance is not an exact eigenstate of H_{QCD} in the strict sense: it is a pole of the S -matrix (or equivalently, of $\log Z$) in the complex energy plane, at $E = m_h - i\Gamma_h/2$. The real part is the mass; the imaginary part is the half-width. In the time domain: the resonance rotates at frequency $m_h c^2/\hbar$ but its amplitude decays exponentially as $e^{-\Gamma_h t/2}$ —the pattern dissolves. The dissolution is the strong decay: the amplitude flows from the resonance’s Fock configuration to the final-state hadrons’ configurations.

In the Fock-Born framework: the resonance is a quasi-stable pattern in the amplitude distribution. The Born observable $\hat{\mathcal{O}}_h$ still defines the resonance (by its quantum numbers), but the projected superposition has a finite lifetime: the pattern rotates and simultaneously leaks into the decay-product eigensubspaces. The width Γ_h is the rate at which the squared length of the projection decreases.

The distinction is quantitative, not qualitative. A “stable” hadron is a pattern with $\Gamma = 0$ (in QCD alone); a resonance is a pattern with $\Gamma > 0$. Both are superpositions of Fock basis vectors; both rotate at frequency $m_h c^2/\hbar$; both are defined by their quantum numbers. The only difference is whether H_{QCD} has matrix elements connecting the hadron’s eigensubspace to lighter-hadron eigensubspaces (if yes: $\Gamma > 0$, the pattern leaks; if no: $\Gamma = 0$, the pattern is locked).

The colour singlet constraint

Not every eigenstate of H_{QCD} is a hadron. Only the colour-singlet eigenstates are physical: the confinement property of QCD ensures that the projected superposition of any physical state onto a colour-non-singlet Fock configuration has exactly zero squared length. Formally:

$$P_{\text{singlet}} |\text{physical}\rangle = |\text{physical}\rangle, \quad P_{\text{non-singlet}} |\text{physical}\rangle = 0, \quad (41.17)$$

where P_{singlet} projects onto the colour-singlet subspace of $\mathcal{F}_{\text{QCD}}$. This is a superselection rule: the physical Hilbert space is $\mathcal{F}_{\text{QCD}}^{\text{singlet}} \subset \mathcal{F}_{\text{QCD}}$.

The colour-singlet constraint determines which combinations of quarks and gluons can form hadrons:

$$q\bar{q} : \quad \mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{8} \oplus \mathbf{1} \rightarrow \text{colour singlet } \mathbf{1} \quad (\text{mesons}), \quad (41.18)$$

$$qqq : \quad \mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{10} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{1} \rightarrow \text{colour singlet } \mathbf{1} \quad (\text{baryons}), \quad (41.19)$$

$$gg : \quad \mathbf{8} \otimes \mathbf{8} = \mathbf{27} \oplus \cdots \oplus \mathbf{1} \rightarrow \text{colour singlet } \mathbf{1} \quad (\text{glueballs}), \quad (41.20)$$

$$qq\bar{q}\bar{q} : \quad \text{various} \rightarrow \mathbf{1} \quad (\text{tetraquarks}), \quad (41.21)$$

$$qqqq\bar{q} : \quad \text{various} \rightarrow \mathbf{1} \quad (\text{pentaquarks}). \quad (41.22)$$

The colour-singlet wavefunctions are specific: $\delta^i_j/\sqrt{3}$ for mesons (colour-anticolour); $\epsilon^{ijk}/\sqrt{6}$ for baryons (fully antisymmetric in colour). Every hadron must have one of these colour structures.

The parton distribution functions: the Fock decomposition measured

The Fock decomposition (41.8) is not merely a theoretical construct: it is directly measured in deep inelastic scattering (DIS).

In DIS, a high-energy virtual photon ($Q^2 \gg \Lambda_{\text{QCD}}^2$) strikes a hadron and resolves its constituents. The cross section is expressed in terms of *parton distribution functions* (PDFs):

$$f_{i/h}(x, Q^2) = \text{probability of finding parton } i \text{ carrying momentum fraction } x \text{ at resolution } Q^2, \quad (41.23)$$

where i runs over all quarks, antiquarks, and gluons, and $x = k^+/P^+$ is the longitudinal momentum fraction (in light-cone coordinates).

The PDFs are the Born rendering of the momentum-fraction observable:

$$f_{i/h}(x, Q^2) = \sum_{\text{Fock sectors}} \int d\Gamma |\phi_h^{(\text{sector})}(\{k_j\})|^2 \delta(x - k_i^+/P^+), \quad (41.24)$$

where $d\Gamma$ denotes the integration over all momenta and quantum numbers except the momentum fraction x of parton i . This is the OR sum: the squared length of the projected superposition onto the eigensubspace with parton i carrying fraction x , summed over all other parton configurations (the degeneracy).

The PDFs satisfy sum rules that are exact consequences of the Fock decomposition:

$$\text{momentum:} \quad \sum_i \int_0^1 dx \, x \, f_{i/h}(x, Q^2) = 1 \quad (\text{all momentum accounted for}), \quad (41.25)$$

$$\text{valence:} \quad \int_0^1 dx \, [f_{u/p}(x) - f_{\bar{u}/p}(x)] = 2 \quad (\text{two valence } u\text{'s in the proton}), \quad (41.26)$$

$$\int_0^1 dx \, [f_{d/p}(x) - f_{\bar{d}/p}(x)] = 1 \quad (\text{one valence } d \text{ in the proton}). \quad (41.27)$$

The momentum sum rule (41.25) says: the total momentum of all partons (quarks, antiquarks, gluons) equals the hadron's momentum. Experimentally, the gluon carries about 50% of the proton's momentum at $Q^2 \sim 10 \text{ GeV}^2$: the gluon Fock sectors are not a small correction but a dominant feature.

The Q^2 dependence of the PDFs—the DGLAP evolution—is the running of the Fock decomposition with resolution:

$$\frac{\partial f_{i/h}(x, Q^2)}{\partial \log Q^2} = \frac{\alpha_s(Q^2)}{2\pi} \sum_j \int_x^1 \frac{dy}{y} P_{i \leftarrow j}\left(\frac{x}{y}\right) f_{j/h}(y, Q^2), \quad (41.28)$$

where $P_{i \leftarrow j}(z)$ are the Altarelli–Parisi splitting functions (the probability for parton j to radiate and produce parton i with fraction z of its momentum). As Q^2 increases (finer resolution), more gluon and sea-quark Fock sectors become visible: the proton’s Fock decomposition unfolds with the resolution of the Born observable.

Summary: hadrons in the Fock-Born framework

Concept	Formal statement
Fock space	$\mathcal{F}_{\text{QCD}} = \bigoplus \mathcal{H}^{(n_q, n_{\bar{q}}, n_g)}$, built from quark and gluon creation operators (41.1)
Hadron	Eigenstate of H_{QCD} in the colour-singlet sector: $H_{\text{QCD}} h\rangle = E_h h\rangle$, with definite J^{PC} , I , S , C , B' , B
Fock decomposition	$ h\rangle = \sum_{\text{sectors}} \phi_h^{(\text{sector})} \prod b^\dagger d^\dagger a^\dagger \Omega\rangle$: a superposition over all Fock sectors (41.8)
Single frequency	The entire superposition rotates at $\omega_h = m_h c^2 / \hbar$ (41.12): H_0 and H_{int} balance exactly
Born observable	$\hat{M}^2 _{\mathcal{Q}, \text{singlet}}$: invariant mass in a specific quantum-number channel (41.14)
Degeneracy	$d_h = \dim \mathcal{H}_{\text{internal}}^{(h)}$: all Fock configs. sharing mass m_h and quantum numbers $\mathcal{Q}$
Colour singlet	$P_{\text{non-singlet}} \text{physical}\rangle = 0$: an exact superselection rule, not a dynamical suppression
Stable hadron	$\Gamma = 0$ in QCD: the pattern rotates without leaking
Resonance	$\Gamma > 0$: the pattern rotates and leaks; pole of $\log Z$ at $E = m_h - i\Gamma/2$
PDFs	The Born rendering of the momentum-fraction observable: $f_{i/h}(x) = \ P_{x,i} h_t\ ^2$, the OR sum over all other parton configurations
DGLAP	The running of the Fock decomposition with the resolution Q^2 of the Born observable

The game engine knows only quarks and gluons: creation operators $b^\dagger$, $d^\dagger$, $a^\dagger$ and the QCD Hamiltonian H_{QCD} . The hadrons are the eigenstates —the stable patterns in the amplitude distribution—that rotate at definite frequencies. The hundreds of named particles (π , K , ρ , ω , p , n , Λ , Σ , Ξ , Ω , Δ , J/ψ , Υ , D , B , $X(3872)$, ...) are the hundreds of eigenfrequencies of this Hamiltonian in the colour-singlet sector. The names label the patterns; the Born observable determines which patterns are visible at which resolution; and the Fock space contains them all, simultaneously, as superpositions of quarks and gluons.

41.3 Representation notation: bold numbers, $\otimes$, and $\oplus$

Throughout the Standard Model sections, we have written expressions like $\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{8} \oplus \mathbf{1}$ and assigned fermions to representations such as $(\mathbf{2}, -1)$ or $(\mathbf{3}, \mathbf{2}, +\frac{1}{3})$. This subsection makes the notation precise, connecting it to the concrete vector spaces and operations already developed in this book.

Bold numbers are dimensions of vector spaces

A bold number $\mathbf{n}$ denotes an irreducible representation of a group G : a vector space V of dimension n on which the group acts, with no proper invariant subspace.

Notation	Group	Vector space	What it means
1	Any	$\mathbb{C}^1$	Singlet: G acts trivially ($g \cdot v = v$)
2	$SU(2)$	$\mathbb{C}^2$	Doublet: the fundamental (spin- $\frac{1}{2}$)
3	$SU(2)$	$\mathbb{C}^3$	Triplet: the adjoint (spin-1)
3	$SU(3)$	$\mathbb{C}^3$	Fundamental: quark colour
$\bar{\mathbf{3}}$	$SU(3)$	$\mathbb{C}^3$	Conjugate fundamental: antiquark colour
8	$SU(3)$	$\mathbb{C}^8$	Adjoint: gluon colour

The bold number is the dimension of the representation, not the group itself. The same number can refer to different representations of different groups (e.g. **3** of $SU(2)$ is the spin-1 representation; **3** of $SU(3)$ is the fundamental quark representation). Context—which group is being discussed—determines the meaning.

The bar notation. A bar $\bar{\mathbf{n}}$ denotes the *conjugate* representation: the group acts by $(g^{-1})^T$ instead of g . For $SU(2)$, the fundamental **2** and its conjugate $\bar{\mathbf{2}}$ are equivalent ($\mathbf{2} \cong \bar{\mathbf{2}}$, related by $i\sigma_2$), so the bar is often omitted. For $SU(3)$, the fundamental **3** and its conjugate $\bar{\mathbf{3}}$ are *not* equivalent: a quark transforms as **3** and an antiquark as $\bar{\mathbf{3}}$. The distinction matters physically: a quark and an antiquark have opposite colour charges.

Connection to the Fock space. A creation operator $b_a^\dagger$ that creates a particle in representation $\mathbf{n}$ carries an index $a \in \{1, 2, \dots, n\}$. Under a group transformation $g \in G$, the creation operator transforms as:

$$b_a^\dagger \rightarrow D(g)_{ab} b_b^\dagger, \quad (41.29)$$

where $D(g)$ is the $n \times n$ matrix representing g in representation $\mathbf{n}$. The Fock basis vector $|a\rangle = b_a^\dagger |0\rangle$ transforms as a vector in the n -dimensional space V : the representation labels how the particle's internal index rotates under the symmetry.

The direct product $\otimes$: combining two representations

The direct product (or tensor product) $\mathbf{m} \otimes \mathbf{n}$ is the representation on the combined space $V_m \otimes V_n$, obtained by taking two particles—one in representation $\mathbf{m}$, one in representation $\mathbf{n}$ —and asking how the pair transforms.

Concretely: if particle A has index $a \in \{1, \dots, m\}$ and particle B has index $b \in \{1, \dots, n\}$, the pair is described by the two-index object T_{ab} , which transforms as:

$$T_{ab} \rightarrow D^{(m)}(g)_{aa'} D^{(n)}(g)_{bb'} T_{a'b'}. \quad (41.30)$$

This is the direct product we constructed concretely in the Fock section: the $m \times n$ table with entry (a, b) , or equivalently a stacked vector of dimension $m \times n$.

The direct product space has dimension $m \times n$, but it is generally *not* irreducible: it decomposes into a direct sum of irreducible representations.

The direct sum $\oplus$: decomposing into irreducible pieces

The direct sum $\mathbf{p} \oplus \mathbf{q}$ is a vector space of dimension $p + q$ that contains $\mathbf{p}$ and $\mathbf{q}$ as non-overlapping, orthogonal subspaces. We constructed this concretely in the Fock section: zero-pad each basis vector into the larger space, with $\mathbf{p}$ occupying the first p entries and $\mathbf{q}$ occupying the last q entries.

A general vector in $\mathbf{p} \oplus \mathbf{q}$ can have amplitude in both subspaces. The group G acts on each subspace independently: it does not mix $\mathbf{p}$ with $\mathbf{q}$. The direct sum is *reducible*: it contains two invariant subspaces.

The decomposition of a tensor product. The key use of the direct sum is to decompose a tensor product into irreducible representations. The tensor product $\mathbf{m} \otimes \mathbf{n}$ is a space of dimension mn , and it decomposes as:

$$\mathbf{m} \otimes \mathbf{n} = \bigoplus_k \mathbf{d}_k, \quad \sum_k d_k = mn. \quad (41.31)$$

The dimensions of the irreducible pieces sum to mn . The decomposition depends on the group G and is determined by the Clebsch–Gordan coefficients—the unitary change-of-basis matrix between the product basis $|a, b\rangle$ and the irreducible basis $|d_k, \alpha\rangle$.

Examples

$SU(2)$: angular momentum addition. The representations of $SU(2)$ are labelled by spin $j = 0, \frac{1}{2}, 1, \frac{3}{2}, \dots$, with dimension $2j + 1$. The tensor product decomposes as:

$$(\mathbf{2j_1} + \mathbf{1}) \otimes (\mathbf{2j_2} + \mathbf{1}) = \bigoplus_{J=|j_1-j_2|}^{j_1+j_2} (\mathbf{2J} + \mathbf{1}). \quad (41.32)$$

For two spin- $\frac{1}{2}$ particles:

$$\mathbf{2} \otimes \mathbf{2} = \mathbf{3} \oplus \mathbf{1}. \quad (41.33)$$

Dimension check: $2 \times 2 = 4 = 3 + 1$. The $\mathbf{3}$ is the spin-1 triplet ($J = 1$, three states $m = -1, 0, +1$); the $\mathbf{1}$ is the spin-0 singlet ($J = 0$, one state $m = 0$). In component language: the symmetric combinations $|\uparrow\uparrow\rangle, (|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle)/\sqrt{2}, |\downarrow\downarrow\rangle$ form the triplet; the antisymmetric combination $(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)/\sqrt{2}$ is the singlet.

This is the addition of angular momenta from the earlier section, expressed in representation notation.

$SU(3)_c$: quark colour. The representations of $SU(3)$ relevant for QCD:

$$\mathbf{3} \otimes \bar{\mathbf{3}} = \mathbf{8} \oplus \mathbf{1} \quad (3 \times 3 = 9 = 8 + 1), \quad (41.34)$$

$$\mathbf{3} \otimes \mathbf{3} = \mathbf{6} \oplus \bar{\mathbf{3}} \quad (3 \times 3 = 9 = 6 + 3), \quad (41.35)$$

$$\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} = \mathbf{10} \oplus \mathbf{8} \oplus \mathbf{8} \oplus \mathbf{1} \quad (3^3 = 27 = 10 + 8 + 8 + 1). \quad (41.36)$$

The physical content:

- **Mesons** ($q\bar{q}$): the $\mathbf{3} \otimes \bar{\mathbf{3}}$ decomposes into an octet $\mathbf{8}$ (the colour structure of a gluon) and a singlet $\mathbf{1}$ (the colour structure of a meson). Physical mesons are colour singlets: they live in the $\mathbf{1}$.
- **Baryons** (qqq): the $\mathbf{3}^{\otimes 3}$ contains a singlet $\mathbf{1}$ —the totally antisymmetric ϵ^{ijk} combination. Physical baryons live in this singlet.
- **Gluons**: carry the adjoint representation $\mathbf{8}$. The eight gluons correspond to the eight generators of $SU(3)$.

The Standard Model quantum numbers. A field in the Standard Model is labelled by its representations under each factor of the gauge group. The notation $(\mathbf{n}_c, \mathbf{n}_L, Y)$ specifies the $SU(3)_c$ representation, the $SU(2)_L$ representation, and the $U(1)_Y$ hypercharge:

Field	$SU(3)_c$	$SU(2)_L$	$U(1)_Y$
Left-handed quark Q_L	3	2	$+\frac{1}{3}$
Right-handed up quark u_R	3	1	$+\frac{4}{3}$
Right-handed down quark d_R	3	1	$-\frac{2}{3}$
Left-handed lepton L_L	1	2	-1
Right-handed electron e_R	1	1	-2
Higgs ϕ	1	2	$+1$
Gluon G_μ^a	8	1	0
W boson W_μ^a	1	3	0
B boson B_μ	1	1	0

A **1** means the field is a singlet under that factor: the symmetry does not act on it. A **3** under $SU(3)_c$ means the field carries a colour index $a \in \{r, g, b\}$. A **2** under $SU(2)_L$ means the field is a doublet with an isospin index $i \in \{1, 2\}$. The full representation of a field under the Standard Model gauge group is the direct product of its representations under each factor:

$$Q_L \sim (\mathbf{3}, \mathbf{2}, +\frac{1}{3}) \implies Q_L^{ai} : a \in \{1, 2, 3\}, i \in \{1, 2\}, \quad (41.37)$$

a field with $3 \times 2 = 6$ internal components (colour $\times$ isospin), plus the continuous indices (momentum, spin) that label the Fock basis.

The notation in the Fock-Born framework

In the Fock-Born language:

- A bold number $\mathbf{n}$ specifies the dimension of the internal index on a creation operator: $b_a^\dagger$ with $a \in \{1, \dots, n\}$.
- The direct product $\mathbf{m} \otimes \mathbf{n}$ is the two-index Fock basis vector $|a, b\rangle$: the $m \times n$ table from the Fock section.
- The direct sum $\mathbf{p} \oplus \mathbf{q}$ is the zero-padded stacking from the Fock section: basis vectors in a $(p + q)$ -dimensional space.
- The decomposition $\mathbf{m} \otimes \mathbf{n} = \bigoplus_k \mathbf{d}_k$ is a change of basis (a unitary transformation) from the product basis $|a, b\rangle$ to the irreducible basis $|d_k, \alpha\rangle$ —exactly the transformation theory of the earlier section.
- The Born rule for a symmetry quantum number projects h_t onto the subspace of a specific irreducible representation and computes the squared length. A colour-singlet measurement projects onto the **1** subspace within $\mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3}$; the squared length is the probability of finding a colour-singlet baryon.

The bold-number notation is a shorthand for the concrete constructions already in the book. Every $\mathbf{n}$ is a vector space built by direct product; every $\oplus$ is a zero-padded stacking; every decomposition is a change of basis. The notation compresses these operations into a single line—like Dirac notation compresses the Hilbert space operations—making the representation-theoretic structure of the Standard Model visible at a glance.

41.4 The Standard Model miracles

The previous subsections catalogued the Standard Model's content: the gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$, the fermion representations, the Higgs doublet, the 19 parameters, the hadron

spectrum. We now step back and ask: *why does this specific content work?*

The Standard Model is not a generic chiral gauge theory. A randomly chosen assignment of fermion representations and hypercharges would almost certainly be inconsistent: the gauge anomalies would not cancel, the theory would not be unitary, and the path integral would not define a consistent quantum theory. The SM fermion content satisfies a set of highly non-trivial consistency conditions that were not imposed by the model-builders but discovered after the fact. These are the “miracles” of the Standard Model — structural coincidences that the formalism requires but does not explain.

Miracle 1: anomaly cancellation

The anomaly cancellation conditions (section ??) require:

$$\sum_{\text{left-handed}} \text{Tr}_R[T^a\{T^b, T^c\}] = \sum_{\text{right-handed}} \text{Tr}_R[T^a\{T^b, T^c\}] \quad (41.38)$$

for all combinations of generators from $SU(3)_c$, $SU(2)_L$, and $U(1)_Y$. There are six independent conditions (listed in the table of section ??). Two are automatic (the purely $SU(3)$ and $SU(2)$ anomalies vanish by group theory). The remaining four are non-trivial: they relate the hypercharges of quarks to those of leptons.

The miracle is that all four conditions are satisfied *simultaneously* by the observed fermion content. The cancellation depends on:

The number of colours $N_c = 3$. The quark contributions to the anomaly are multiplied by N_c (each quark comes in N_c colours). The $[U(1)_Y]^3$ anomaly cancels only because $N_c = 3$ produces the correct numerical coefficient. For $N_c = 2$ or $N_c = 4$: the anomaly does not cancel with the observed hypercharge assignments.

The quark-lepton pairing. Each generation contains both quarks and leptons, with hypercharges that are correlated: $Y_{Q_L} = +1/6$, $Y_{u_R} = +2/3$, $Y_{d_R} = -1/3$ for quarks; $Y_{L_L} = -1/2$, $Y_{e_R} = -1$ for leptons. The anomaly cancels generation by generation: the quark contribution from one generation is cancelled by the lepton contribution from the same generation. If quarks and leptons had unrelated hypercharges, the cancellation would fail.

The fractional charges. The quarks have electric charges $Q_u = +2/3$ and $Q_d = -1/3$ — fractions of the electron charge. These are not arbitrary: they are the unique values (given $N_c = 3$) that make the anomaly cancel.

Miracle 2: charge quantisation from anomaly cancellation

In a $U(1)$ gauge theory, the charge can be any real number: there is no group-theoretic reason for quantisation. This is unlike $SU(N)$, where the representations are labelled by integers and the charges are automatically quantised. The hypercharge Y is a $U(1)_Y$ charge, and *a priori* could take any value.

The anomaly cancellation conditions change this. Treat the hypercharges as unknowns: $Y_{Q_L} = q$, $Y_{u_R} = u$, $Y_{d_R} = d$, $Y_{L_L} = \ell$, $Y_{e_R} = e$, and impose the four non-trivial anomaly conditions. The system:

$$[U(1)_Y]^3 : \quad N_c \cdot 2q^3 + 2\ell^3 - N_c u^3 - N_c d^3 - e^3 = 0, \quad (41.39)$$

$$[SU(3)]^2[U(1)] : \quad 2q - u - d = 0, \quad (41.40)$$

$$[SU(2)]^2[U(1)] : \quad N_c q + \ell = 0, \quad (41.41)$$

$$\text{grav} \times [U(1)] : \quad N_c \cdot 2q + 2\ell - N_c u - N_c d - e = 0. \quad (41.42)$$

This is a system of four equations in five unknowns. One overall normalisation is free (the $U(1)$ charge normalisation). The solution, up to this normalisation:

$$q = \frac{1}{6}, \quad u = \frac{2}{3}, \quad d = -\frac{1}{3}, \quad \ell = -\frac{1}{2}, \quad e = -1 \quad (\text{with } N_c = 3). \quad (41.43)$$

These are precisely the observed Standard Model hypercharges.

The miracle: the hypercharges are *not free parameters*. They are determined (up to normalisation) by the requirement that the theory be anomaly-free. Charge quantisation — the fact that all observed charges are integer multiples of $e/3$ — is a consequence of gauge consistency, not an additional postulate.

This is sometimes taken as evidence for grand unification: if the SM gauge group is embedded in a simple group like $SU(5)$ or $SO(10)$, then all charges are automatically quantised (simple groups have only integer-labelled representations). But the anomaly argument shows that charge quantisation follows from the SM alone, without GUTs: the consistency of the chiral gauge theory forces the charges to be rational.

Miracle 3: three generations

The Standard Model has three generations (families) of fermions: (e, ν_e, u, d) , (μ, ν_μ, c, s) , (τ, ν_τ, t, b) . Each generation has the same gauge quantum numbers; the generations differ only in their Yukawa couplings (masses). The number of generations $N_g = 3$ is not determined by gauge consistency: the anomaly cancellation works for any N_g (it cancels generation by generation).

Three facts constrain N_g :

Asymptotic freedom requires $N_g \leq 5$. The QCD beta function $b_3 = 11N_c/3 - 2N_f/3$ is negative (asymptotic freedom) only if $N_f < 11N_c/2$. Each generation contributes $N_f = 2$ quark flavours, so $N_g < 11N_c/4 = 33/4 = 8.25$. Actually, at the two-loop level and including threshold corrections, asymptotic freedom is lost for $N_g \geq 6$ (where $N_f = 12$, close to the Banks–Zaks fixed point). So $N_g \leq 5$ for QCD to confine.

CP violation requires $N_g \geq 3$. The CKM matrix for N_g generations is an $N_g \times N_g$ unitary matrix. The number of physical CP-violating phases:

$$n_{\text{CP}} = \frac{(N_g - 1)(N_g - 2)}{2}. \quad (41.44)$$

For $N_g = 2$: $n_{\text{CP}} = 0$. No CP violation is possible with two generations. For $N_g = 3$: $n_{\text{CP}} = 1$. One physical phase δ in the CKM matrix — the Kobayashi–Maskawa phase, for which they received the 2008 Nobel Prize. The observed CP violation in K and B meson decays requires $N_g \geq 3$.

The Z width measures N_g directly. The total Z decay width $\Gamma_Z = \Gamma_{\text{had}} + 3\Gamma_\ell + N_\nu\Gamma_\nu$, measured at LEP: $N_\nu = 2.984 \pm 0.008$. This is a Born-layer measurement: the event rate at the Z peak determines the number of invisible decay channels, hence the number of light neutrino species, hence $N_g = 3$ (assuming one neutrino per generation).

Baryogenesis may require $N_g \geq 3$. One of Sakharov’s three conditions for generating the matter-antimatter asymmetry is CP violation. With $N_g = 2$: no CKM phase, no CP violation in the quark sector, and (within the SM) no baryogenesis. The observed baryon asymmetry $\eta_B \sim 6 \times 10^{-10}$ requires at least the single CKM phase of $N_g = 3$. (The SM CP violation is quantitatively too small for the observed η_B , but the qualitative requirement $N_g \geq 3$ stands.)

Summary: $N_g = 3$ is the minimum. Three is the smallest number of generations that allows CP violation in the quark sector. It is consistent with asymptotic freedom. It is measured directly from the Z width. Whether it is “explained” by a deeper principle (a family symmetry, a compactification geometry in string theory, or something else) is an open question.

The interconnections

The three miracles are not independent. They form a web of consistency conditions that constrain the SM content far more tightly than the gauge group alone:

Input	Condition	Consequence
Gauge group G_{SM}	Anomaly cancellation	Hypercharges fixed (up to normalisation)
Anomaly cancellation	Requires $N_c = 3$	Fractional quark charges $Q = 2/3, -1/3$
Charge quantisation	$Q = I_3 + Y/2$	Integer lepton charges, fractional quark
CP violation	Requires $N_g \geq 3$	CKM phase $\delta \neq 0$
Asymptotic freedom	Requires $N_g \leq 5$	Confinement, hadrons
Z width	Measures $N_\nu = 3$	$N_g = 3$ confirmed
$G_{\text{SM}} + N_c = 3 + N_g = 3$	All conditions satisfied	The Standard Model

The Standard Model is not a free choice of representations. It is one of a small number of chiral gauge theories (with the gauge group $SU(3) \times SU(2) \times U(1)$) that are simultaneously anomaly-free, asymptotically free, and CP-violating. The specific fermion content — three generations of quarks and leptons with the observed hypercharges — is largely *dictated* by these consistency conditions, not freely chosen.

The Fock-Born reading

The three miracles have a unified Fock-Born interpretation: they are *consistency conditions on the engine*.

Anomaly cancellation is the condition that the Fock-layer path integral $\int \mathcal{D}\Phi e^{iS_{\text{SM}}}$ defines a consistent quantum theory. If the anomaly does not cancel, the gauge symmetry is broken at the quantum level, the Ward–Takahashi identities fail, the S -matrix is not unitary, and the Born-layer probabilities do not sum to one. Anomaly cancellation is not a property of the Born-layer rendering; it is a property of the Fock-layer engine’s internal consistency.

Charge quantisation is a consequence of the engine’s consistency: the requirement that the path integral be gauge-invariant constrains the $U(1)_Y$ charges to be rational numbers. The Born-layer observable — the measured electric charge of the electron, the proton, the pion — is quantised because the engine’s gauge symmetry demands it.

Three generations is the minimum N_g for which the engine can produce CP violation. CP violation is a Born-layer observable (measured in K and B meson decays), but the engine must have at least three generations of fermion creation operators for the CKM matrix to have a physical phase. The engine’s particle content determines what the Born layer can render.

The Standard Model miracles are not aesthetic coincidences. They are the conditions under which the Fock-layer engine is self-consistent and capable of producing the Born-layer outputs (unitarity, charge quantisation, CP violation) that are observed. A randomly chosen chiral gauge

theory would fail one or more of these conditions and would not define a consistent quantum theory. The observed Standard Model is one of the few that passes all tests.

What the miracles do not explain. The miracles constrain the SM content but do not determine it uniquely. The gauge group $SU(3) \times SU(2) \times U(1)$ is an input, not a consequence. The number $N_c = 3$ is measured, not derived. The number $N_g = 3$ is the minimum for CP violation, but $N_g = 4$ or $N_g = 5$ would also be consistent (and are excluded only by the Z width measurement, a Born-layer observation, not a theoretical argument). The 19 parameters (masses, couplings, mixing angles) are completely free within the anomaly-free framework: the miracles determine the *structure* of the SM but not its *parameters*.

Whether the gauge group, the number of colours, and the number of generations are themselves consequences of a deeper structure (grand unification, string compactification, some unknown principle) is the central open question of fundamental physics. The miracles show that the SM is more constrained than it appears; they do not show that it is unique.

41.5 Confinement

The most dramatic non-perturbative consequence of $SU(3)_c$ gauge theory is confinement: quarks and gluons are never observed in isolation. Every Born-layer observable is a colour singlet. The proton (a baryon), the pion (a meson), and the glueball (a bound state of gluons) are all colour singlets; a single quark in the **3** representation is not.

The Wilson loop criterion. Confinement is diagnosed by the Wilson loop, the gauge-invariant observable:

$$W[\mathcal{C}] = \frac{1}{N} \text{Tr} \mathcal{P} \exp \left(i g_s \oint_{\mathcal{C}} T^a G_\mu^a dx^\mu \right), \quad (41.45)$$

where $\mathcal{C}$ is a closed spacetime loop and $\mathcal{P}$ denotes path ordering. The vacuum expectation value $\langle W[\mathcal{C}] \rangle$ probes the static quark-antiquark potential: for a rectangular loop of spatial extent R and temporal extent T :

$$\langle W[R, T] \rangle \sim e^{-V(R)T} \quad (\text{for } T \gg R), \quad (41.46)$$

where $V(R)$ is the static potential between a quark and an antiquark separated by distance R .

Area law ($V(R) \sim \sigma R$): if $\langle W \rangle \sim \exp(-\sigma \cdot \text{Area}[\mathcal{C}])$, the potential is linear. The force $F = -dV/dR = -\sigma$ is constant: it costs an energy $\sigma \cdot \Delta R$ to separate the quarks by an additional distance ΔR . The constant σ is the *string tension* (energy per unit length of the colour flux tube). This is confinement.

Perimeter law ($V(R) \sim c/R$): if $\langle W \rangle \sim \exp(-c \cdot \text{Perimeter}[\mathcal{C}])$, the potential is Coulomb-like. This is deconfinement (the charges are screened).

Lattice QCD confirms the area law for $SU(3)$: at distances $R \gtrsim 0.5 \text{ fm}$, $V(R) = \sigma R - \pi/(12R) + \text{const}$, with $\sigma \approx (440 \text{ MeV})^2 \approx 0.18 \text{ GeV}^2 \approx 0.9 \text{ GeV/fm}$.

The static potential: three regimes.

$$V(R) = \begin{cases} -\frac{4\alpha_s(R)}{3R} & R \ll \Lambda_{\text{QCD}}^{-1} \quad (\text{Coulomb: asymptotic freedom}) \\ \sigma R - \frac{\pi}{12R} & R \gg \Lambda_{\text{QCD}}^{-1} \quad (\text{linear: confinement}) \\ \text{screened} & R \gg 1 \text{ fm} \quad (\text{string breaking: } q\bar{q} \text{ pair creation}) \end{cases} \quad (41.47)$$

At short distances: the one-gluon exchange gives a Coulomb potential with running coupling $\alpha_s(R)$ (the factor $4/3$ is the $SU(3)$ colour Casimir C_F for the fundamental representation). At

intermediate distances: the colour flux collimates into a tube and the potential becomes linear. At very large distances: the flux tube has enough energy to create a $q\bar{q}$ pair from the vacuum (string breaking), and the potential flattens — the original quark-antiquark pair separates into two colour-singlet mesons.

41.6 The colour flux tube

Between a static quark and antiquark at large separation, the colour electric field is confined to a narrow tube of cross-sectional area $A \sim 1/\sigma$. This is qualitatively different from QED, where the electric field spreads as $1/r^2$ (the Coulomb field).

The dual superconductor picture. 't Hooft and Mandelstam proposed that the QCD vacuum is a “dual superconductor”: a condensate of magnetic monopoles that confines colour-electric flux, just as a conventional superconductor (a condensate of Cooper pairs) confines magnetic flux into Abrikosov vortices.

In a Type-II superconductor: a magnetic flux line is confined to a tube of radius $\sim \lambda_L$ (the London penetration depth), with energy per unit length $\sigma_{\text{mag}} \sim (\Phi_0/\lambda_L)^2$. In the dual picture: a colour-electric flux line between quarks is confined to a tube of radius $\sim \Lambda_{\text{QCD}}^{-1} \sim 0.5 \text{ fm}$, with energy per unit length $\sigma \sim \Lambda_{\text{QCD}}^2$.

The dual superconductor picture is not a derivation of confinement (it assumes the magnetic monopole condensate, which has not been rigorously established from the QCD Lagrangian), but it provides the correct qualitative physics: the vacuum expels colour-electric flux, concentrating it into tubes.

Lattice evidence. Lattice QCD directly visualises the flux tube. The chromoelectric field $\langle E_z^a(x_\perp) \rangle$ between a static $q\bar{q}$ pair (computed on the lattice using the plaquette operator in the presence of a Wilson line) shows a Gaussian profile in the transverse direction:

$$\langle E^2(x_\perp) \rangle \propto \exp\left(-\frac{x_\perp^2}{2w^2}\right), \quad w \approx 0.35 \text{ fm}, \quad (41.48)$$

where w is the flux-tube width. The tube width grows logarithmically with the quark separation: $w^2(R) \sim w_0^2 + c \ln(R/R_0)$ (“roughening”), which is a prediction of the effective string theory (see below).

41.7 Glueballs

Gluons carry colour charge and interact with each other. This opens a possibility absent in QED: bound states of pure gauge field with no quarks. These are *glueballs* — colour-singlet states composed entirely of gluons.

The spectrum. Glueballs are classified by J^{PC} (spin, parity, charge conjugation). Lattice QCD predicts the lightest glueball states in pure $SU(3)$ gauge theory (without quarks):

J^{PC}	Mass (MeV)	Status
0^{++}	$1710 \pm 50 \pm 80$	Lightest; scalar
2^{++}	$2390 \pm 30 \pm 120$	Tensor
0^{-+}	$2560 \pm 35 \pm 120$	Pseudoscalar

In full QCD (with dynamical quarks): glueballs mix with $q\bar{q}$ mesons of the same J^{PC} , making experimental identification difficult. The $f_0(1710)$ is the strongest glueball candidate: its

production pattern (enhanced in gluon-rich processes like J/ψ radiative decay, suppressed in $\gamma\gamma$ fusion) is consistent with a dominant glueball component.

The Fock-Born reading. Glueballs are certified collective modes of the pure gauge field. In the Fock space $\mathcal{F}_a^{\text{QCD}}$: the glueball creation operator is a composite of gluon creation operators $a_{\mathbf{k}}^{a\dagger}$, schematically $G^\dagger \sim \sum_{\mathbf{k},a,b} c_{\mathbf{k}}^{ab} a_{\mathbf{k}}^{a\dagger} a_{-\mathbf{k}}^{b\dagger}$ (for 0^{++}). The glueball is certified by a sharp pole in the two-gluon correlator: the spectral function has a peak at $m_G \approx 1710 \text{ MeV}$ with width $\Gamma_G \sim 100\text{--}200 \text{ MeV}$.

In the Fock-Born language: the glueball is a Born-to-Fock promotion of a composite gauge-field observable (a colour-singlet combination of gluon operators) to a Fock-layer degree of freedom (a creation operator in the hadronic Fock space $\mathcal{F}_{\text{hadron}}$). The promotion is less sharp than for the pion (the glueball width is $\sim 10\%$ of its mass), but the spectral peak is clear on the lattice.

41.8 The effective string theory of the flux tube

The colour flux tube between a quark and antiquark is a one-dimensional extended object: a string. Its low-energy dynamics is described by an effective string theory — a two-dimensional field theory on the string worldsheet. This is not an analogy: it is a rigorous effective field theory, derived by integrating out the massive QCD modes and retaining only the transverse fluctuations of the flux tube.

The degrees of freedom. The flux tube in D spacetime dimensions has $D - 2$ transverse oscillation modes: the displacement fields $X^i(\sigma, \tau)$ ($i = 1, \dots, D - 2$), where σ parametrises the position along the string and τ is the worldsheet time. At long wavelengths (wavelengths $\gg w \sim \Lambda_{\text{QCD}}^{-1}$, the tube width), these are the only light degrees of freedom: all other excitations (internal modes of the flux tube, glueball-like excitations of the tube cross section) are massive with masses $\sim \Lambda_{\text{QCD}}$.

The Nambu–Goto action as leading term. The effective action for the transverse fluctuations is organised by the number of derivatives (Weinberg’s power counting applied to the worldsheet). The leading term is the Nambu–Goto action:

$$S_{\text{NG}} = -\sigma \int d^2\xi \sqrt{-\det(\partial_a X^\mu \partial_b X_\mu)}, \quad (41.49)$$

where σ is the string tension ($\sigma \approx 0.18 \text{ GeV}^2$ for QCD) and $\xi^a = (\tau, \sigma)$ are the worldsheet coordinates.

Expanding around a static straight-string configuration $X^0 = \tau$, $X^{D-1} = \sigma$, $X^i = \text{const}$ (for $i = 1, \dots, D - 2$), and writing the transverse fluctuations as $X^i(\sigma, \tau) = h^i(\sigma, \tau)$:

$$S_{\text{NG}} = -\sigma \int d^2\xi \sqrt{1 - (\dot{h}^i)^2 + (h^{i'})^2 - (\dot{h}^i h^{j'} - \dot{h}^j h^{i'})^2 + \dots}, \quad (41.50)$$

where $\dot{h} = \partial_\tau h$ and $h' = \partial_\sigma h$. Expanding the square root to quadratic order:

$$S \approx -\sigma LT + \frac{\sigma}{2} \int d^2\xi [(\dot{h}^i)^2 - (h^{i'})^2] + O(h^4), \quad (41.51)$$

where LT is the worldsheet area (the static string energy $\sigma L \cdot T$). The quadratic action is a free massless two-dimensional field theory: $D - 2$ free scalar fields $h^i(\sigma, \tau)$ on the worldsheet.

The worldsheet CFT. The free-field action (41.51) defines a two-dimensional conformal field theory (CFT₂) with central charge:

$$c = D - 2. \quad (41.52)$$

Each transverse mode contributes $c = 1$ (one free boson). The worldsheet CFT governs the universal long-distance properties of the flux tube: the spectrum of string excitations, the width growth, and the inter-quark potential corrections.

The Lüscher term. The most important prediction of the worldsheet CFT is the Lüscher correction to the static potential. For a string of length R with Dirichlet boundary conditions (fixed endpoints at the quark and antiquark), the zero-point energy of the transverse oscillations:

$$E_0(R) = \sigma R + \frac{(D-2)}{2} \sum_{n=1}^{\infty} \frac{n\pi}{R} = \sigma R - \frac{(D-2)\pi}{24R} \quad (\text{after zeta regularisation}). \quad (41.53)$$

The regularised sum uses $\sum_{n=1}^{\infty} n = \zeta(-1) = -1/12$ (the same zeta regularisation as in the Casimir effect and in the critical dimension of the bosonic string). The result:

$$\boxed{V(R) = \sigma R - \frac{(D-2)\pi}{24R} + O(1/R^3)}. \quad (41.54)$$

The $-\pi/12R$ correction (for $D = 4$: $D - 2 = 2$ transverse modes) is the *Lüscher term*. It is a universal prediction of the effective string theory: it depends only on the number of transverse modes, not on the details of the flux tube (the profile, the internal structure, the coupling constant).

Lattice QCD confirms the Lüscher term to high precision: $V(R) = \sigma R - \pi/(12R) + \text{const}$ for $R \gtrsim 0.5 \text{ fm}$, with the coefficient $-\pi/12$ matching the prediction for $D - 2 = 2$ transverse modes within $\sim 1\%$ accuracy.

Correction terms and the critical dimension. The Nambu–Goto action is not the complete effective action: higher- derivative corrections appear, organised by powers of $1/(\sigma R^2)$ (the ratio of the string scale to the separation):

$$S_{\text{eff}} = S_{\text{NG}} + \int d^2\xi \sqrt{-\gamma} \left[b_2 R_{\text{ws}} + c_1 (\partial^2 h^i)^2 / \sigma + c_2 (K_{ab} K^{ab})^2 / \sigma + \dots \right], \quad (41.55)$$

where γ_{ab} is the induced metric, R_{ws} is the worldsheet Ricci scalar, and K_{ab} is the extrinsic curvature.

A remarkable result (Polchinski and Strominger 1991, Lüscher and Weisz 2004): the first correction beyond Nambu–Goto is constrained by Lorentz invariance. In D spacetime dimensions, the Nambu–Goto string at the quantum level has a Weyl anomaly (a conformal anomaly on the worldsheet) proportional to $D - 26$:

$$\langle T^a_a \rangle_{\text{ws}} \propto \frac{D-26}{12} R_{\text{ws}}. \quad (41.56)$$

At $D = 26$: the anomaly vanishes and the Nambu–Goto string is quantum-mechanically consistent as a fundamental theory (the bosonic string critical dimension). At $D \neq 26$: the anomaly must be cancelled by the leading correction term in (41.55). The Polchinski–Strominger term:

$$S_{\text{PS}} = \frac{26-D}{24\pi} \int d^2\xi \frac{(\partial_+ X \cdot \partial_+^2 X)(\partial_- X \cdot \partial_-^2 X)}{(\partial_+ X \cdot \partial_- X)^2}, \quad (41.57)$$

is uniquely determined by the requirement that Lorentz invariance be restored at the quantum level. Its coefficient is proportional to $26 - D$: it vanishes at the critical dimension and is required for $D < 26$.

The physical consequence: the static potential receives corrections beyond the Lüscher term:

$$V(R) = \sigma R - \frac{(D-2)\pi}{24R} + \frac{c_3}{R^3} + \frac{c_5}{R^5} + \dots, \quad (41.58)$$

where the coefficients c_3, c_5 depend on D through the Polchinski–Strominger correction. Lattice QCD measurements of c_3 are consistent with the Nambu–Goto prediction up to corrections of order $1/(\sigma R^2)^2$ — confirming that the effective string theory description works quantitatively for the QCD flux tube.

The Fock-Born reading. The effective string theory of the flux tube is a nested Fock-Born structure:

Level 1 (the microscopic engine): the QCD Hamiltonian $H_{\text{QCD}}(a)$ on the lattice, with quarks and gluons as Fock-layer degrees of freedom.

Level 2 (the emergent engine): the flux tube, a certified collective mode of the gluon field. Its effective Fock space is the Hilbert space of the worldsheet CFT_2 , with transverse oscillations h^i as the Fock-layer degrees of freedom and the Nambu–Goto action as the effective Hamiltonian.

Level 3 (the Born-layer output): the static potential $V(R) = \sigma R - \pi/(12R) + \dots$, measured on the lattice by computing the Wilson loop $\langle W[R, T] \rangle$.

The Lüscher term $-\pi/(12R)$ is a Born-layer observable of the worldsheet engine, predicted by the worldsheet CFT without any input from the microscopic QCD Hamiltonian. It is a universal prediction: any confining gauge theory in $D = 4$ produces the same $-\pi/(12R)$ correction. This is universality (section ??) at work: the irrelevant operators (the flux tube internal structure, the coupling constant, the quark masses) wash out, and the Born-layer output depends only on the symmetry ($D - 2$ transverse modes) and the single relevant parameter (σ).

41.9 Hadron scattering and Regge trajectories

The connection between the QCD flux tube and string theory is not merely an analogy: it is the historical origin of string theory. The subject was born from hadron scattering data in the 1960s, a decade before QCD was established.

The experimental pattern. In the 1960s, experimentalists observed a striking regularity in the hadron spectrum: mesons and baryons with the same flavour quantum numbers but increasing spin J lie on straight lines when plotted as J vs. M^2 (mass squared):

$$\boxed{J = \alpha_0 + \alpha' M^2}, \quad (41.59)$$

where $\alpha' \approx 0.9 \text{ GeV}^{-2}$ is the universal *Regge slope* and α_0 is an intercept that depends on the quantum numbers. These straight lines are *Regge trajectories*.

Examples (for the ρ meson trajectory):

Hadron	J^{PC}	M (MeV)	M^2 (GeV ²)
$\rho(770)$	1^{--}	775	0.60
$a_2(1320)$	2^{++}	1318	1.74
$\rho_3(1690)$	3^{--}	1689	2.85
$a_4(2040)$	4^{++}	2018	4.07

Plotting these points gives a straight line with slope $\alpha' \approx 0.88 \text{ GeV}^{-2}$.

The rotating string model. The Regge trajectory has a simple mechanical explanation: a relativistic string of tension σ rotating rigidly about its centre.

Consider a string of length $2R$ rotating with angular velocity ω about its midpoint. The endpoints move at the speed of light: $v_{\text{end}} = \omega R = c$, so $\omega = c/R$. The energy and angular momentum of

the rotating string:

$$E = 2 \int_0^R \frac{\sigma dr}{\sqrt{1 - v^2(r)/c^2}} = 2 \int_0^R \frac{\sigma dr}{\sqrt{1 - r^2/R^2}} = \pi \sigma R, \quad (41.60)$$

$$J = 2 \int_0^R \frac{\sigma v(r) r dr}{c^2 \sqrt{1 - v^2(r)/c^2}} = 2 \int_0^R \frac{\sigma \omega r^2 dr}{c^2 \sqrt{1 - r^2/R^2}} = \frac{\pi \sigma R^2}{2c}. \quad (41.61)$$

Eliminating R :

$$J = \frac{E^2}{2\pi\sigma c} = \frac{1}{2\pi\sigma} M^2 c^2 = \alpha' M^2, \quad \alpha' = \frac{1}{2\pi\sigma}. \quad (41.62)$$

The Regge slope is inversely proportional to the string tension: $\alpha' = 1/(2\pi\sigma)$. With $\sigma \approx (440 \text{ MeV})^2 \approx 0.19 \text{ GeV}^2$: $\alpha' \approx 1/(2\pi \times 0.19) \approx 0.84 \text{ GeV}^{-2}$, in good agreement with the observed $\alpha' \approx 0.9 \text{ GeV}^{-2}$.

The Regge trajectory is the spectrum of a rotating relativistic string. Higher-spin hadrons are longer strings rotating faster.

The Veneziano amplitude. Veneziano (1968) wrote down a scattering amplitude that satisfies crossing symmetry (invariance under $s \leftrightarrow t$ channel exchange) and has poles on the Regge trajectories:

$$A(s, t) = \frac{\Gamma(-\alpha(s)) \Gamma(-\alpha(t))}{\Gamma(-\alpha(s) - \alpha(t))}, \quad \alpha(s) = \alpha_0 + \alpha' s, \quad (41.63)$$

where s and t are the Mandelstam variables and Γ is the Euler gamma function. The poles of $\Gamma(-\alpha(s))$ at $\alpha(s) = 0, 1, 2, \dots$ correspond to resonances on the Regge trajectory.

Nambu, Nielsen, and Susskind (1970) independently realised that the Veneziano amplitude is the scattering amplitude of open strings: the poles are the excitation levels of the string, and the crossing symmetry is a consequence of the worldsheet's conformal symmetry. This was the birth of string theory — not as a theory of quantum gravity, but as a theory of hadron scattering.

The Regge tower. A string has infinitely many excitation levels, each with a specific mass and spin. The open string spectrum:

$$M^2 = \frac{1}{\alpha'} (N - a), \quad J_{\max} = N, \quad (41.64)$$

where $N = 0, 1, 2, \dots$ is the oscillator level (the total number of string quanta), a is the normal-ordering constant ($a = 1$ for the open bosonic string, $a = 1/2$ for the open superstring), and $J_{\max}$ is the maximum spin at level N . The leading Regge trajectory connects the states of maximum spin at each level: $J = N = \alpha' M^2 + a$.

The tower of states at each level N is called the Regge tower (or string tower). At level N : the degeneracy grows as $d_N \sim N^{-a} \exp(c\sqrt{N})$ for some constants a and c (the Hagedorn growth), producing exponentially many states at high excitation.

The historical reversal. The original string theory of hadrons was abandoned in the 1970s when QCD was established as the correct theory of the strong interaction. QCD explains confinement, asymptotic freedom, and the hadron spectrum from first principles. The string description of hadrons was demoted from a fundamental theory to an effective description of the flux tube.

But the string theory survived by changing its domain: Scherk and Schwarz (1974) proposed that the spin-2 massless state of the closed string (present in all string theories but absent from the hadron spectrum) should be identified with the graviton. The string tension was re-interpreted

as $\sigma \sim M_P^2 \sim 10^{38} \text{ GeV}^2$ (instead of $\sigma \sim \Lambda_{\text{QCD}}^2 \sim 0.04 \text{ GeV}^2$): string theory became a theory of quantum gravity rather than of hadrons.

The Fock-Born reading. The Regge trajectory is the Born-layer spectrum of the QCD flux tube — the rendering of the worldsheet engine’s excitation levels. Each hadron on the trajectory is a certified collective mode: a specific rotation/vibration state of the string, with quantum numbers (J, M) that match the rotating-string prediction.

The Veneziano amplitude is the Born-layer scattering cross section computed from the worldsheet engine. It was the first amplitude computed from a string theory, and its agreement with hadron scattering data was the first evidence that the QCD flux tube behaves as a string.

The effective string theory of the flux tube (section 41.8) is the modern, rigorous version of this story: the Nambu–Goto action plus corrections, derived as an effective field theory of the QCD flux tube, with the Lüscher term and the Polchinski–Strominger correction confirmed by lattice QCD. The hadronic string is alive, not as a fundamental theory, but as a certified emergent description of the QCD vacuum.

41.10 Lattice QCD: the engine made explicit

Every non-perturbative result quoted in the previous subsections — the string tension, the static potential, the Lüscher term, the flux tube profile, the glueball masses, the hadron spectrum — comes from lattice QCD. This is the Fock-Born framework’s most complete realisation: a quantum field theory defined directly on a lattice, with no reference to a continuum limit, producing Born-layer predictions that agree with experiment.

We describe the construction in detail, because lattice QCD is not merely a computational technique applied to QCD. It *is* QCD — the only non-perturbative definition of the theory.

Wilson’s lattice gauge theory

Wilson (1974) gave the first non-perturbative definition of a gauge theory by placing it on a hypercubic spacetime lattice with spacing a . The key insight: the gauge field is not a vector $G_\mu^a(x)$ living at the sites, but a group element $U_\mu(x)$ living on the *links* between sites.

Link variables. Each link $(x, x + a\hat{\mu})$ carries a *link variable*:

$$U_\mu(x) = \mathcal{P} \exp \left(i g_s \int_x^{x+a\hat{\mu}} T^a G_\nu^a dy^\nu \right) \in SU(3), \quad (41.65)$$

which is the parallel transporter (Wilson line) from site x to its neighbour $x + a\hat{\mu}$. On the lattice: $U_\mu(x)$ is a 3×3 unitary matrix with unit determinant, one per link. There is no need to define the continuum gauge field G_μ^a ; the link variable $U_\mu(x)$ is the fundamental object.

Under a gauge transformation $\Omega(x) \in SU(3)$ at each site:

$$U_\mu(x) \rightarrow \Omega(x) U_\mu(x) \Omega^\dagger(x + a\hat{\mu}). \quad (41.66)$$

The link variable transforms by left-multiplication at the source and right-multiplication at the target. This is exact gauge invariance, not an approximation: the lattice gauge symmetry is $SU(3)^{N_{\text{sites}}}$, the same infinite-dimensional group as in the continuum (one $SU(3)$ per site instead of per spacetime point).

The plaquette. The smallest gauge-invariant object built from link variables is the *plaquette*: the trace of the product of four links around an elementary square:

$$U_{\mu\nu}(x) = U_\mu(x) U_\nu(x + a\hat{\mu}) U_\mu^\dagger(x + a\hat{\nu}) U_\nu^\dagger(x). \quad (41.67)$$

The plaquette is gauge-invariant because the Ω 's cancel around the closed loop: $\Omega(x)$ from the first link cancels $\Omega^\dagger(x)$ from the last.

The Wilson gauge action. The lattice gauge action:

$$S_G[U] = \frac{2}{g_s^2} \sum_x \sum_{\mu < \nu} \text{Re Tr} [\mathbf{1} - U_{\mu\nu}(x)]. \quad (41.68)$$

This is a sum over all plaquettes, each contributing a term proportional to how much $U_{\mu\nu}$ deviates from the identity. In the limit $a \rightarrow 0$: expanding $U_\mu(x) \approx \mathbf{1} + ig_s a T^a G_\mu^a + O(a^2)$, the plaquette becomes $U_{\mu\nu} \approx \mathbf{1} + ig_s a^2 T^a G_{\mu\nu}^a - \frac{g_s^2 a^4}{2} T^a T^b G_{\mu\nu}^a G_{\mu\nu}^b + \dots$, and the action reduces to:

$$S_G \rightarrow \frac{1}{2g_s^2} \sum_x a^4 \text{Tr}(G_{\mu\nu}^a G^{a\mu\nu}) = \frac{1}{4} \int d^4x G_{\mu\nu}^a G^{a\mu\nu}, \quad (41.69)$$

which is the continuum Yang–Mills action. The lattice action (41.68) is the Yang–Mills action: its value on smooth, slowly varying configurations agrees with the continuum expression up to corrections of order a^2 .

The key property. Wilson's action (41.68) maintains exact $SU(3)$ gauge invariance at every lattice spacing a . This is unlike Lorentz invariance (which is broken to the cubic group and must be rendered) or diffeomorphism invariance (which is broken completely and must be rendered). The gauge symmetry is not an emergent, rendered symmetry: it is exact on the lattice. The Wald identity (section 29.6) holds exactly at every a , and the gluon masslessness is protected at every scale.

Fermions on the lattice

Putting quarks on the lattice is harder than putting gluons. The naïve lattice Dirac operator:

$$D_{\text{naive}}(x, y) = \sum_\mu \gamma^\mu \frac{U_\mu(x) \delta_{y, x+\hat{\mu}} - U_\mu^\dagger(y) \delta_{y, x-\hat{\mu}}}{2a} + m \delta_{x, y} \quad (41.70)$$

has a problem: the *fermion doubling*. In the free case ($U_\mu = \mathbf{1}$), the lattice dispersion relation $\omega(k) = \sin(ka)/a$ has zeros at both $k = 0$ and $k = \pi/a$ (the corners of the Brillouin zone). Each corner produces an independent fermion species. In $d = 4$: $2^4 = 16$ doublers per flavour—15 too many.

This is not a technical inconvenience but a theorem: the Nielsen–Ninomiya no-go theorem (1981) states that any lattice fermion action that is local, translational-invariant, and has the correct continuum limit must have an equal number of left- and right-handed fermion species. Chiral fermions (with unequal numbers of left- and right-handed modes) cannot be placed on the lattice without sacrificing one of these properties.

Wilson fermions. Wilson's solution: add a second-derivative term $\frac{a}{2} \bar{\psi} \nabla^2 \psi$ that gives the doublers a mass $\sim 1/a$ (they decouple at low energies):

$$D_W = D_{\text{naive}} - \frac{a}{2} \sum_\mu \nabla_\mu^* \nabla_\mu. \quad (41.71)$$

The Wilson term breaks chiral symmetry explicitly (it is a mass-like term), so the chiral limit $m \rightarrow 0$ requires fine-tuning of the bare mass (an additive renormalization). The rendering-stability condition determines the critical bare mass $m_c(g_s)$ at which the physical pion mass vanishes.

Staggered fermions. Kogut and Susskind’s alternative: distribute the four spinor components of the Dirac field across different lattice sites, reducing the doubling from 16 to 4. The remaining four “tastes” become degenerate in the continuum limit. Each physical quark flavour is represented by taking the fourth root of the fermion determinant: $\det(D_{\text{stag}})^{1/4}$. The validity of this “rooting” procedure is theoretically controversial but empirically successful.

Domain wall and overlap fermions. The modern solution: extend the lattice to five dimensions. In the domain wall formulation (Kaplan 1992): the physical fermion is a zero mode bound to a four-dimensional domain wall in a five-dimensional lattice. The chiral symmetry is preserved exponentially well (the chiral symmetry breaking is $\sim e^{-L_5/a}$, where L_5 is the extent of the fifth dimension). The overlap formulation (Neuberger 1998) is the $L_5 \rightarrow \infty$ limit, giving exact lattice chiral symmetry satisfying the Ginsparg–Wilson relation $\gamma_5 D + D \gamma_5 = a D \gamma_5 D$.

The Euclidean path integral on the lattice

The lattice QCD partition function:

$$Z = \int \prod_{x,\mu} dU_\mu(x) \prod_{x,f} d\bar{\psi}_f(x) d\psi_f(x) e^{-S_G[U] - \sum_f \bar{\psi}_f D[U] \psi_f}, \quad (41.72)$$

where dU_μ is the Haar measure on $SU(3)$ (the unique left-and-right-invariant measure on the group), the product over f runs over quark flavours, and $D[U]$ is the lattice Dirac operator (Wilson, staggered, or overlap) coupled to the gauge field through the link variables.

Integrating out the Grassmann-valued quark fields analytically:

$$Z = \int \prod_{x,\mu} dU_\mu(x) \prod_f \det D_f[U] e^{-S_G[U]}. \quad (41.73)$$

The fermion determinant $\det D_f[U]$ encodes all quark-loop effects. In the *quenched approximation*: $\det D_f \rightarrow 1$ (no dynamical quarks; quark loops are neglected). This is computationally cheap but unphysical: it misses string breaking, sea-quark effects, and the η' mass. Modern lattice QCD includes the full fermion determinant (“dynamical fermions”).

Monte Carlo evaluation. The lattice integral (41.73) is evaluated by importance sampling: generate gauge configurations $\{U_\mu(x)\}$ with probability proportional to $\prod_f \det D_f[U] \cdot e^{-S_G[U]}$, and estimate expectation values as sample averages (section 23.5). The algorithms (Hybrid Monte Carlo for dynamical fermions, heatbath for quenched) generate a Markov chain of configurations that converges to the Boltzmann distribution.

A typical modern lattice QCD calculation uses: $N_s^3 \times N_t = 64^3 \times 128$ lattice sites, lattice spacing $a \approx 0.06\text{--}0.12$ fm, physical volume $L \approx 4\text{--}6$ fm, $N_f = 2 + 1$ or $2 + 1 + 1$ dynamical quark flavours (up, down, strange, and optionally charm), and $O(10^3\text{--}10^4)$ statistically independent gauge configurations. Each configuration is a collection of $4 \times 64^3 \times 128 \approx 1.3 \times 10^9$ $SU(3)$ matrices — a dataset of $\sim 10^{10}$ real numbers.

Physical results

Lattice QCD produces the most precise non-perturbative predictions in quantum field theory:

The hadron spectrum. The masses of the light hadrons (π , K , N , Δ , Σ , Ξ , Ω , ρ , etc.), computed from Euclidean correlators $C(\tau) = \langle \mathcal{O}(\tau) \mathcal{O}^\dagger(0) \rangle_E$ and extracting the exponential decay rate $m = -\lim_{\tau \rightarrow \infty} \partial_\tau \ln C(\tau)$, agree with experiment to 1–3% accuracy. The BMW collaboration (2008) achieved sub-percent agreement for the proton and neutron masses, with

all sources of systematic uncertainty (finite volume, lattice spacing, quark mass extrapolation) controlled.

This is a Born-layer prediction from first principles: the input is the QCD Lagrangian with m_u , m_d , m_s , and α_s (four parameters); the output is the entire hadron spectrum (~ 50 measured states). No model, no approximation, no adjustable parameters beyond the four inputs.

The strong coupling. $\alpha_s(m_Z) = 0.1184 \pm 0.0012$ from lattice QCD (FLAG 2021 average), in excellent agreement with the world average from all methods ($\alpha_s(m_Z) = 0.1179 \pm 0.0009$).

CKM matrix elements. The CKM elements $|V_{us}|$, $|V_{cb}|$, $|V_{ub}|$ require non-perturbative hadronic matrix elements (form factors for semileptonic decays $B \rightarrow \pi \ell \nu$, $B \rightarrow D^* \ell \nu$, $K \rightarrow \pi \ell \nu$) that can only be computed by lattice QCD. The lattice results, combined with experimental decay rates, determine the CKM elements to $\sim 1\%$ precision.

The QCD phase diagram. At finite temperature ($T > 0$): the lattice finds a crossover (not a sharp phase transition) from the hadronic phase to the quark-gluon plasma at $T_c \approx 155$ MeV. The transition temperature, the equation of state $p(T)$, and the trace anomaly $(\epsilon - 3p)/T^4$ are all computed on the lattice and compared to heavy-ion collision data from RHIC and the LHC.

The string tension and glueball masses. $\sigma = (440 \pm 10 \text{ MeV})^2$ from the Wilson loop. $m_{0^{++}} = 1710 \pm 50 \pm 80 \text{ MeV}$ from glueball correlators. The Lüscher term coefficient: $-\pi/(12R)$ confirmed to 1%.

The lattice-first viewpoint: vindication by practice

Lattice QCD is the existence proof that the lattice-first viewpoint of this book is not merely philosophical — it is how the most precise non-perturbative calculations in physics are actually performed.

The lattice is the definition. The continuum QCD Lagrangian $\mathcal{L}_{\text{QCD}}$ does not define a quantum theory: the path integral $\int \mathcal{D}G \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{iS_{\text{QCD}}}$ is a formal expression that requires regularisation. The lattice provides the regularisation — but in the lattice-first viewpoint, it is more than that. The lattice *is* the definition of QCD: the partition function (41.72) at finite a is a well-defined, finite-dimensional integral. It does not need a continuum limit to be meaningful.

The continuum limit $a \rightarrow 0$ is a computational convenience: it simplifies the algebra (continuum Feynman rules are easier than lattice Feynman rules) and makes Lorentz invariance manifest. But the physics is already present at finite a : the hadron masses, the string tension, the CKM elements, and the QCD phase diagram are all computed at finite a and extrapolated to $a \rightarrow 0$ only to remove the $O(a^2)$ lattice artefacts (the irrelevant operators of section ??).

Gauge invariance is exact; Lorentz invariance is rendered. Wilson's construction maintains exact $SU(3)$ gauge invariance at every a . The link variables are group elements; the plaquette action is built from gauge-invariant traces; the fermion determinant is gauge-invariant by construction. No gauge-fixing is needed. The Wald identity holds at every a . The gluon mass is exactly zero at every a .

Lorentz invariance, by contrast, is broken to the cubic group O_h on a hypercubic lattice. The lattice dispersion relation $\omega(\mathbf{k})$ deviates from $\omega = |\mathbf{k}|$ at $|\mathbf{k}| \sim \pi/a$. These deviations are lattice artefacts — irrelevant operators suppressed by $O(a^2 k^2)$ — that vanish at long wavelengths. Lorentz invariance is a rendered symmetry: it holds to accuracy $O(a/\lambda)^2$ for Born-layer observables at wavelength $\lambda \gg a$.

This distinction — exact gauge symmetry, rendered Lorentz symmetry — is the central structural insight of the lattice-first viewpoint, visible in every lattice QCD calculation.

The rendering-stability condition in practice. A lattice QCD calculation at spacing a_1 gives hadron masses $m_H(a_1)$. A second calculation at spacing $a_2 < a_1$ gives $m_H(a_2)$. The rendering-stability condition: $m_H(a_1) = m_H(a_2) + O(a^2)$. In practice:

$$m_H(a) = m_H^{\text{phys}} + c_2 a^2 + c_4 a^4 + \cdots, \quad (41.74)$$

and the physical mass m_H^{phys} is obtained by fitting multiple lattice spacings and extrapolating to $a = 0$. This is the rendering-stability condition of section ??, applied to a real computation.

The extrapolation to $a = 0$ removes the irrelevant operators ($O(a^2)$ lattice artefacts) and isolates the Born-layer output. But the extrapolation is a refinement, not a necessity: the results at finite $a \sim 0.06$ fm already agree with experiment to $\sim 1\%$. The lattice artefacts are small because the lattice spacing is much smaller than the hadronic scale: $a \ll 1/\Lambda_{\text{QCD}} \sim 1$ fm.

The sign problem as a Fock-layer limitation. When the fermion determinant $\det D_f[U]$ is not positive (e.g., at finite baryon chemical potential $\mu_B > 0$, or with a topological θ -term), the weight $\det D \cdot e^{-S_G}$ is not a probability distribution and Monte Carlo sampling fails. This is the *sign problem*: a computational limitation of the Euclidean path integral, not a physical one.

The sign problem does not mean the physics is undefined. The Hamiltonian $H_{\text{QCD}}(a)$ is well-defined at any μ_B and θ : it is a Hermitian operator on the Fock space $\mathcal{F}_a^{\text{QCD}}$. The spectrum, the ground state, and all Born-layer observables exist. What fails is one specific *method* for computing them (Euclidean Monte Carlo). The Hamiltonian formulation (section ??) does not have a sign problem: the time evolution e^{-iHt} is always unitary, and the spectrum of H is always real.

This is a concrete illustration of the two-language principle (section 21.11): the Hamiltonian and the path integral compute the same physics, but they have different computational domains. The path integral excels when $e^{-S_E} \geq 0$ (Monte Carlo works); the Hamiltonian excels when it does not.

The lattice-first thesis, restated. Lattice QCD vindicates the central claim of this book:

Lattice-first thesis	Lattice QCD evidence
The lattice is the engine, not a regulariser	Z_{QCD} at finite a is the definition
Gauge symmetry is exact on the lattice	Wilson’s link-variable construction
Lorentz invariance is rendered	$O(a^2)$ artefacts, confirmed and removed
The continuum limit is a convenience	Results at finite a agree with experiment
The rendering-stability condition works	$m_H(a) = m_H^{\text{phys}} + O(a^2)$, confirmed
Born-layer observables are a -independent	Hadron masses, α_s , CKM: same at all a
The Hamiltonian side has no sign problem	H_{QCD} is well-defined at all μ_B

The lattice-first viewpoint is not a philosophical preference. It is the viewpoint from which every non-perturbative QCD prediction is actually computed. The proton mass, the pion decay constant, the CKM matrix elements, and the QCD phase diagram are all computed on a lattice, at finite a , with exact gauge invariance and rendered Lorentz invariance. The continuum limit is the last step, not the first.

The rest of physics — the Standard Model, general relativity, the effective string theory of the flux tube — is conjectured to have the same structure. The lattice is the Fock layer; the

continuum physics is the Born-layer rendering. Lattice QCD is the proof that this architecture works.

41.11 The hierarchy problem: what it is and what it is not

The hierarchy problem is traditionally stated as: the one-loop correction to the Higgs mass is $\delta m_H^2 \sim 1/a^2$ (“quadratically divergent”), so the bare mass and the correction must cancel to 30 decimal places to produce $m_{H,\text{phys}} = 125 \text{ GeV}$. This statement is perturbative, and as we now argue, the apparent cancellation is largely an artefact of perturbation theory. The genuine non-perturbative content is more subtle and more interesting.

The perturbative artefact

The one-loop formula:

$$m_{H,\text{phys}}^2 = m_H^2(\text{bare}) + \frac{c\pi^2}{a^2} + O(\lambda \ln(1/a)), \quad c = \frac{-6y_t^2 + \frac{9}{4}g^2 + \frac{3}{4}g'^2 + 6\lambda}{16\pi^2}, \quad (41.75)$$

presents the physical mass as the *sum* of two large numbers: $m_H^2(\text{bare}) \sim -10^{34} \text{ GeV}^2$ and $c\pi^2/a^2 \sim +10^{34} \text{ GeV}^2$, yielding $(125)^2 \text{ GeV}^2$. The “30-digit cancellation” is between these two terms.

But this cancellation is a feature of the perturbative bookkeeping, not of the physics.

The e^{-100} analogy. Consider computing e^{-100} by its Taylor series:

$$e^{-100} = \sum_{n=0}^{\infty} \frac{(-100)^n}{n!} = 1 - 100 + 5000 - 166667 + \dots \approx 3.7 \times 10^{-44}. \quad (41.76)$$

The individual terms are enormous: $|(-100)^{42}/42!| \approx 10^{40}$. They alternate in sign and “cancel” to 44 decimal places. A naive reading would declare this a “fine-tuning problem”: terms of order 10^{40} must cancel to produce a result of order 10^{-44} . The function e^{-100} is “unnaturally small” relative to the individual terms of its Taylor expansion.

But this is absurd. The function e^{-100} is a single, well-defined number. No cancellation is involved in its *existence*. The cancellation is an artefact of expanding a smooth function in a power series around a point ($x = 0$) that is far from the region of interest ($x = -100$). A different expansion (e.g., around $x = -100$) would show no large terms and no cancellation.

The perturbative hierarchy problem has exactly this structure. The “bare mass” and “one-loop correction” are two terms in a Taylor expansion (in the coupling constants λ, y_t, g, g') of a single quantity: $m_{H,\text{phys}}^2$. The terms are individually large ($\sim 1/a^2$) because the expansion is organised by powers of the couplings, not by the physical mass. The cancellation between them is an artefact of this organisation.

The non-perturbative computation

On the lattice, non-perturbatively: there is no expansion, no “bare mass,” no “one-loop correction,” and no cancellation.

The lattice Hamiltonian $H(a)$ is defined with a set of bare parameters $\theta(a) = (m_H^2(\text{bare}), \lambda, y_t, g, g', \dots)$. The physical Higgs mass is extracted from the two-point correlator:

$$\langle H(\mathbf{x}, t) H(\mathbf{0}, 0) \rangle \xrightarrow{|\mathbf{x}| \rightarrow \infty} A e^{-m_{H,\text{phys}} |\mathbf{x}|}. \quad (41.77)$$

One number in ($m_H^2(\text{bare})$), one number out ($m_{H,\text{phys}}$). The relationship between them is a function:

$$m_{H,\text{phys}} = f(m_H^2(\text{bare}), \lambda, y_t, g, g', a), \quad (41.78)$$

computed by Monte Carlo or by diagonalising $H(a)$. The function f is perfectly smooth. No individual terms are large. No cancellation occurs. The “hierarchy problem” as a cancellation between large numbers is invisible in the non-perturbative computation.

The perturbative expansion (41.75) is a Taylor expansion of the function f in powers of λ , y_t^2 , etc. The individual Taylor coefficients are of order $1/a^2$ because f depends on a through the lattice cutoff. But the function f itself is whatever it is: a smooth, finite, well-defined map from bare parameters to physical observables. The “30-digit cancellation” is between Taylor coefficients, not between physical quantities.

The non-perturbative sensitivity: the phase diagram

The perturbative cancellation is an artefact. But the non-perturbative physics does contain a genuine sensitivity, of a different kind: the proximity to a phase transition.

The lattice Higgs theory in $d = 4$ has a phase diagram in the space of bare couplings ($m_H^2(\text{bare}), \lambda$). There is a *critical surface* separating the symmetric phase ($\langle\phi\rangle = 0$) from the broken phase ($\langle\phi\rangle \neq 0$). The physical Higgs mass is controlled by the distance to this critical surface:

$$m_{H,\text{phys}} \sim \frac{1}{a} |m_H^2(\text{bare}) - m_{H,c}^2(\lambda, a)|^\nu, \quad (41.79)$$

where $m_{H,c}^2$ is the critical value (the bare mass at which the correlation length diverges) and ν is the correlation length exponent.

For the Higgs sector in $d = 4$ (which is in the universality class of the Gaussian fixed point, since ϕ^4 is trivial in $d = 4$): $\nu = 1/2$, so:

$$m_{H,\text{phys}}^2 \sim \frac{1}{a^2} |m_H^2(\text{bare}) - m_{H,c}^2|. \quad (41.80)$$

A small physical mass $m_{H,\text{phys}} \cdot a \ll 1$ requires the bare mass to be close to the critical value:

$$|m_H^2(\text{bare}) - m_{H,c}^2| \sim m_{H,\text{phys}}^2 \cdot a^2 \sim (125 \text{ GeV})^2 \cdot \ell_P^2 \sim 10^{-30}. \quad (41.81)$$

The bare mass must be within 10^{-30} of the critical value.

This is NOT the same as the perturbative cancellation. The perturbative problem says: two large numbers cancel. The non-perturbative problem says: the operating point is very close to a critical surface. The distinction matters:

The perturbative version involves large terms in an expansion. It disappears if you don’t expand — it is a property of the Taylor series, not of the function.

The non-perturbative version involves the phase-diagram location of the engine. It does not disappear — it is a property of the function f near the critical surface, regardless of how you compute f .

But the non-perturbative sensitivity has a different flavour. The question “why is the system near-critical?” is a familiar question in statistical mechanics. Near-critical systems are ubiquitous (water near the critical point, magnets near T_c , liquid crystals, superfluids). Sometimes the proximity to criticality has a dynamical explanation (self-organised criticality, cosmological relaxation). Sometimes it is a consequence of the continuum limit itself: taking $m_{H,\text{phys}} \cdot a \ll 1$ (the rendered Higgs mass is much smaller than the lattice scale) *requires* the bare mass to be near the critical surface, by the definition (41.79) of the physical mass. The “fine-tuning” is the statement that the continuum limit exists, not a sickness.

What does “protection” protect against?

The traditional statement: “chiral symmetry protects the fermion mass; gauge symmetry protects the gauge boson mass; no symmetry protects the Higgs mass.” What does “protects” mean, precisely?

In perturbation theory: “Protection” means the large $1/a^2$ terms in the perturbative expansion are absent.

For the photon: the Ward identity forces $\Pi^{\mu\nu} = (q^2\eta^{\mu\nu} - q^\mu q^\nu)\Pi(q^2)$ — no $\eta^{\mu\nu}$ piece. The one-loop correction to m_γ^2 is zero: $\delta m_\gamma^2 = 0$, not just “small.” The dangerous $1/a^2$ term is absent from the expansion.

For a fermion: chiral symmetry forces $\delta m \propto m$. The correction is multiplicative: $m_{\text{phys}} = m[1 + c \ln(1/a) + \dots]$. The dangerous $1/a$ term (which would appear without chiral symmetry) is absent. Only logarithms remain.

For the Higgs: no symmetry removes the $1/a^2$ term. The expansion has $\delta m_H^2 \propto 1/a^2$ at every loop order.

In this perturbative reading: “protection” means the Taylor coefficients of f are well-behaved (no powers of $1/a$ in the expansion). This is a property of the perturbative expansion, not necessarily of the function f itself.

Non-perturbatively: “Protection” means the distance to the critical surface is controlled by a symmetry.

For the photon: gauge invariance is exact on the lattice (section 41.10). The photon mass $m_\gamma = 0$ is exact at every a , because the Ward identity holds at every a . The critical surface for photon mass generation does not exist in the phase diagram: there is no phase with $m_\gamma \neq 0$ within the gauge-invariant theory.

For a fermion with $m = 0$: if chiral symmetry is exact, the critical surface for fermion mass generation is at $m_{\text{bare}} = 0$ exactly. No tuning is needed: $m_{\text{bare}} = 0$ is the natural point, fixed by the symmetry. The correlation length exponent is irrelevant because the system sits exactly on the critical surface by symmetry.

For the Higgs: no symmetry fixes the bare mass to any particular value. The critical surface exists at $m_{H,c}^2(\lambda, a) \neq 0$, and the distance to it is a free parameter. The proximity (41.81) is a fact about where the engine operates, not a consequence of its architecture.

So the non-perturbative meaning of “protection” is: *a symmetry fixes the operating point to sit on (or near) the critical surface, so no tuning is required.* The unprotected Higgs requires its operating point to be chosen near the critical surface without any symmetry reason.

Particle	Perturbative protection	Non-perturbative protection	Status
Photon	$\delta m^2 = 0$ (Ward identity)	No $m_\gamma \neq 0$ phase exists	Exact (gauge)
Fermion	$\delta m \propto m$ (chiral)	$m_c = 0$ by chiral symmetry	Exact (at $m = 0$)
Higgs	$\delta m^2 \propto 1/a^2$ (no protection)	$m_c^2 \neq 0$; proximity is a free param.	Unprotected
Pion	$\delta m^2 \propto m_q$ (chiral)	Near-critical by approximate symmetry	Approximate

The Fock-Born summary

Three statements, clearly separated:

1. The perturbative hierarchy problem is an artefact. The “30-digit cancellation” between $m_H^2(\text{bare})$ and $c\pi^2/a^2$ is a cancellation between Taylor coefficients of a smooth function f . It is analogous to computing e^{-100} by its Taylor series: the terms are large but the function is not. On the lattice, non-perturbatively, there is no expansion, no large terms, and no visible cancellation. The perturbative hierarchy problem is self-inflicted.

2. The non-perturbative sensitivity is real but different. The physical Higgs mass is controlled by the distance to a critical surface: $m_{H,\text{phys}} \sim a^{-1}|m_H^2(\text{bare}) - m_{H,c}^2|^{1/2}$. A small physical mass requires the bare mass to be close to the critical value. This proximity is a non-perturbative fact, independent of any expansion. But it is a fact about the phase diagram, not about cancellations. The question “why near-critical?” replaces the question “why do large terms cancel?”

3. “Protection” fixes the operating point. A symmetry “protects” a mass by fixing the operating point to sit on the critical surface (gauge: the critical surface doesn’t exist; chiral: $m_c = 0$ is the symmetric point). Without protection: the operating point is a free parameter, and its proximity to the critical surface is a fact about the engine’s calibration. This proximity may have a dynamical explanation (SUSY, composite Higgs, cosmological relaxation) or may simply be a boundary condition on the engine.

The Born-layer observer sees $m_{H,\text{phys}} = 125 \text{ GeV}$ and moves on. The perturbative cancellation, the phase-diagram proximity, and the question of protection are all Fock-layer concerns about the engine’s internal structure. The rendering is stable regardless.

Large numbers are not crises

The “30-digit cancellation” is the number $m_H^2 \cdot a^2 \sim 10^{-30}$ (if $a \sim \ell_P$), presented as a crisis. But in lattice units this number simply says: the Higgs correlation length is $\xi_H = 1/m_H \sim 10^{15}$ lattice spacings.

This is a large number, but it is the same *kind* of large number as $N_A \approx 6 \times 10^{23}$ (atoms in a mole), $N_{\text{baryon}} \sim 10^{57}$ (baryons in a star), or the number of voxels in a three-dimensional lattice simulation. It characterises the scale separation in the problem: how many lattice spacings fit inside the physical correlation length. A model with many degrees of freedom naturally has parameters that are large in lattice units. The bare mass $m_H^2(\text{bare}) \sim 1/a^2$ is $O(1)$ in lattice units — it is a lattice-scale number, measuring something about the lattice, like the number of plaquettes per unit area.

Nobody finds it “unnatural” that the phonon wavelength in a crystal ($\sim \text{mm}$) is 10^7 times the atomic spacing ($\sim \text{\AA}$): the phonon “mass” in lattice units is $m_{\text{phonon}} \cdot a \sim 10^{-7}$. Nobody calls this a hierarchy problem. It is what it means to have a low-energy excitation on a fine lattice.

The Higgs is the same, with a larger exponent: $m_H \cdot a \sim 10^{-15}$ instead of 10^{-7} . The exponent is bigger because the lattice is finer (Planck scale rather than atomic scale) and the excitation is heavier (125 GeV rather than meV). But the structure is identical: a macroscopic correlation length on a microscopic lattice. The number 10^{-30} (or 10^{-15} , or 10^{-7}) is the ratio of two physical scales, not a symptom of disease.

The perturbative framing converts this innocent ratio into a “cancellation” by writing $m_{H,\text{phys}}^2 = m_H^2(\text{bare}) + c/a^2$: two large numbers differing by a small amount. But the two numbers are not independent physical quantities that happen to nearly cancel. They are two terms in a

Taylor expansion of a single smooth function f , and their near-equality is an artefact of the expansion (the e^{-100} effect of section 41.11). On the lattice, non-perturbatively: there is one number in ($m_H^2(\text{bare})$), one number out ($m_{H,\text{phys}}$), and a smooth function connecting them. No cancellation. Just a big lattice.

A remark on continuum thinking

The hierarchy problem has shaped the direction of fundamental physics for four decades. It motivated supersymmetry at the TeV scale, composite Higgs models, extra dimensions, and ultimately billions of dollars of collider investment. The superpartners were not found.

The lattice-first viewpoint suggests a reason: the problem, as traditionally stated, is an artefact of continuum perturbation theory, and the solutions were designed to cure the artefact.

The pattern is older than the hierarchy problem. In the 1930s, Dirac struggled with the infinite self-energy of the electron in QED: $E_{\text{self}} = \int_0^\infty dk \, k^2/(k^2 + m^2) = \infty$. The infinity is a consequence of treating the electron as a point particle in a continuum — of integrating the electromagnetic field to arbitrarily short distances. Dirac believed the infinity signalled a fundamental defect in the theory and spent decades searching for a reformulation. But the infinity is an artefact of $\mathbb{R}^3$: on a lattice, $E_{\text{self}} = \sum_{|k| < \pi/a} k^2/(k^2 + m^2) < \infty$. The theory was never sick; the continuum made it look sick.

Renormalization (Schwinger, Tomonaga, Feynman) tamed the infinities by absorbing them into the physical parameters. This works — spectacularly so ($g - 2$ of the electron to 12 decimal places) — but it left a residue of discomfort: the procedure subtracts infinity from infinity. Dirac never accepted it. From the lattice-first viewpoint, there was nothing to accept: on the lattice, every quantity is finite, the “subtraction” is a finite reparametrisation, and the discomfort dissolves.

The hierarchy problem is the same story, one generation later. The continuum perturbative expansion produces $\delta m_H^2 \sim 1/a^2$, which looks like a cancellation crisis. SUSY was introduced to cancel the $1/a^2$ terms — to make the Taylor coefficients of f well-behaved. But the function f was never ill-behaved: it is a smooth, finite map from bare parameters to physical observables. The “crisis” was in the expansion, not in the function.

The cosmological constant problem follows the same pattern. The continuum calculation gives $\rho_\Lambda^{\text{QFT}} \sim 1/a^4$ (the sum of zero-point energies to infinite momentum), and the comparison $\rho_\Lambda^{\text{obs}}/\rho_\Lambda^{\text{QFT}} \sim 10^{-123}$ is declared the “worst prediction in physics.” But if gravity is emergent (section 43.13), Λ is a Born-layer output (a_0 coefficient), not a bare parameter. There is no bare Λ_0 to cancel against. The “ 10^{-123} ” is the ratio of a measured number to a perturbative continuum artefact.

In each case — the self-energy infinity, the hierarchy problem, the cosmological constant — the continuum formulation produces a crisis (an infinity, a cancellation, a mismatch), the crisis motivates a search for new physics, and the lattice reveals that the crisis was in the formulation. The continuum is an extraordinarily powerful computational tool, but it is not physically fundamental. Mistaking the tool for the physics leads to mistaking artefacts of the tool for problems of the physics.

The genuine open questions — why is the Higgs light? why three generations? why $\alpha \approx 1/137$? — are questions about the spectrum and structure of the lattice Hamiltonian $H(a)$. They are the same kind of question as “why is $m_e/m_t \sim 10^{-5}$?”: spectral questions about the engine’s boundary conditions. They do not involve infinities, cancellations, or fine-tunings. They are hard because we do not know the engine well enough, not because the engine is sick.

41.12 The mass gap problem: what should be proved

The Clay Mathematics Institute’s Millennium Prize Problem on “Yang–Mills Existence and Mass Gap” is often cited as the most important open problem at the interface of mathematics and physics. We argue here that the problem, as formulated, asks the wrong question from the standpoint of physics — and that the physically correct question is both more natural and closer to being answered.

The Clay formulation

The official problem statement (Jaffe and Witten, 2000) asks for two things:

Existence. Prove that for any compact simple gauge group G , a quantum Yang–Mills theory exists on $\mathbb{R}^4$ and satisfies the Wightman axioms (or the equivalent Osterwalder–Schrader axioms for the Euclidean theory): a Hilbert space $\mathcal{H}$, a unitary representation of the Poincaré group, a unique vacuum state, a set of operator-valued distributions (the fields) satisfying locality and spectral conditions, and a well-defined scattering theory.

Mass gap. Prove that the theory has a mass gap $\Delta > 0$: the spectrum of the Hamiltonian H satisfies $\text{spec}(H) \subseteq \{0\} \cup [\Delta, \infty)$, with $\Delta > 0$ being the mass of the lightest particle.

Both requirements refer to a theory defined on $\mathbb{R}^4$: a continuum theory in infinite volume with exact Lorentz invariance. No lattice, no finite volume, no lattice spacing.

What the problem actually asks

The Clay problem asks whether the Born-layer rendering (the continuum QFT) can be promoted to a self-standing mathematical object, independent of the engine that produces it. In the Fock-Born language: it asks whether the screen image can exist without the game engine.

This is a question about the mathematical status of the continuum limit, not about the physics of Yang–Mills theory. The physics —the spectrum, the masses, the scattering amplitudes, the confinement mechanism—is already present on the lattice at finite a . The Clay problem asks whether the $a \rightarrow 0$ limit of the lattice theory exists as a rigorous mathematical object satisfying the Wightman axioms.

The Wightman axioms themselves encode assumptions about the rendering:

Axiom 1 (Hilbert space and Poincaré invariance): the theory lives on a Hilbert space carrying a continuous unitary representation of the Poincaré group. On the lattice: the Hilbert space exists, but the Poincaré group is broken to the Euclidean group $O_h \times \mathbb{Z}$ (the cubic group times discrete translations). Full Poincaré invariance is a rendered symmetry that holds at $\lambda \gg a$.

Axiom 2 (spectral condition): the energy-momentum spectrum lies in the forward light cone. On the lattice: the spectrum of $H(a)$ is bounded below and has a mass gap (established numerically), but the dispersion relation $\omega(\mathbf{k})$ deviates from $\sqrt{|\mathbf{k}|^2 + m^2}$ at $|\mathbf{k}| \sim \pi/a$. The spectral condition is satisfied at $\lambda \gg a$ and violated at $\lambda \sim a$.

Axiom 3 (locality): fields at spacelike separation commute. On the lattice: fields at different sites commute (or anticommute for fermions) exactly. Locality is actually *better* on the lattice (exact at the site level) than in the continuum (where it is an axiom about distributional fields).

Axiom 4 (completeness): the vacuum is cyclic for the field algebra. On the lattice: this holds for the lattice field algebra on the finite lattice.

The axioms that fail on the lattice (Axioms 1 and 2) are precisely the ones that encode Poincaré invariance—a rendered symmetry. The axioms that hold on the lattice (Axioms 3 and 4) are the ones that encode the physical content.

The lattice-first reformulation

From the lattice-first viewpoint, the physically correct question is:

Does the lattice Yang–Mills Hamiltonian $H(a)$ have a mass gap $\Delta(a) > 0$ for all $a > 0$, and is the gap stable under the rendering-stability condition: $\Delta(a) = \Delta_{\text{phys}} + O(a^2)$ as $a/\lambda \rightarrow 0$?

This reformulation has three components:

Component 1: at each fixed lattice spacing a , with N_s^3 spatial sites, the Hamiltonian $H(a)$ is a well-defined Hermitian operator on a finite-dimensional Hilbert space (for the gauge sector; infinite-dimensional but separable for the full gauge+matter sector). Its spectrum is discrete. The mass gap $\Delta(a) = E_1(a) - E_0(a)$ (the energy of the lightest excitation above the vacuum) is a well-defined number for each a .

Component 2: the gap must be positive: $\Delta(a) > 0$ for all $a > 0$. This is the statement that the theory confines (the lightest colour-singlet excitation has nonzero mass) at every lattice spacing.

Component 3: the gap must be rendering-stable: $\Delta(a_1) \approx \Delta(a_2)$ for $a_1, a_2 \ll \lambda$ (where λ is the physical wavelength of the gap excitation). This is the requirement that the gap is a property of the physics (a Born-layer observable), not of the lattice spacing (a Fock-layer parameter).

None of these components requires a continuum limit. None requires the Wightman axioms. None requires Poincaré invariance. They require only that the lattice Hamiltonian has a spectral gap and that the gap is insensitive to the lattice spacing.

What lattice QCD already knows

The lattice evidence for the mass gap is overwhelming:

Numerical computation of the gap. The mass gap is computed from the exponential decay of the Euclidean two-point correlator of a colour-singlet operator $\mathcal{O}$ (e.g., the plaquette for the 0^{++} glueball):

$$C(\tau) = \langle \mathcal{O}(\tau) \mathcal{O}^\dagger(0) \rangle_E \xrightarrow{\tau \rightarrow \infty} |\langle 1 | \mathcal{O} | 0 \rangle|^2 e^{-\Delta \tau}. \quad (41.82)$$

The gap Δ is extracted from the exponential decay rate. For pure $SU(3)$ gauge theory: $\Delta = m_{0^{++}} \approx 1710 \pm 50 \text{ MeV}$ (the scalar glueball mass).

Rendering stability of the gap. The gap has been computed at multiple lattice spacings: $a = 0.05, 0.08, 0.10, 0.12 \text{ fm}$, etc. The results:

$$m_{0^{++}}(a) = m_{0^{++}}^{\text{phys}} + c_2 a^2 \sigma + O(a^4), \quad (41.83)$$

where c_2 is a lattice-artefact coefficient and σ is the string tension (used to set the scale). The extrapolation to $a \rightarrow 0$ is smooth: no sign of the gap vanishing or diverging. The rendering-stability condition is satisfied empirically to the precision of the computation ($\sim 3\text{--}5\%$).

The string tension confirms confinement. The Wilson loop expectation value $\langle W[R, T] \rangle \sim e^{-\sigma R T}$ (area law) is confirmed at all lattice spacings, with $\sigma(a) = \sigma_{\text{phys}} + O(a^2)$. The string tension is rendering-stable: confinement persists at all a .

The transfer matrix is positive. On the lattice, the Euclidean partition function Z_E can be written as $Z_E = \text{Tr}(T^{N_t})$, where T is the transfer matrix (the Euclidean time-evolution operator for one lattice step). For the Wilson gauge action: T is a positive self-adjoint operator on the lattice Hilbert space ($T > 0$), which guarantees reflection positivity (the Osterwalder–Schrader condition for reconstructing a Hilbert space with a positive inner product). The existence of $T > 0$ is *proven*, not assumed: it follows from the positivity of the Wilson gauge action’s Boltzmann weight.

What remains to be proved

The gap between the lattice evidence and a rigorous proof is:

The thermodynamic limit. The lattice calculations are performed at finite volume ($L = N_s \cdot a$). The spectral gap $\Delta(a, L)$ at finite volume is rigorously positive (the Hamiltonian on a finite lattice has discrete spectrum with a unique ground state, by the Perron–Frobenius theorem applied to $T > 0$). What must be shown is that the gap survives the infinite-volume limit: $\lim_{L \rightarrow \infty} \Delta(a, L) = \Delta(a) > 0$.

This is a statement about the lattice model, not about the continuum. It is the lattice analogue of demonstrating that a statistical mechanical model in the thermodynamic limit retains a spectral gap (the mass gap of the correlation length: $\xi^{-1} = \Delta > 0$). For many lattice models (e.g., the 2D Ising model above T_c), this is proven. For lattice Yang–Mills in $d = 4$: it is established numerically but not yet rigorously.

The rendering-stability condition. The functions $\Delta(a)$, $\sigma(a)$, $m_{0++}(a)$ must be shown to satisfy $\Delta(a) = \Delta_{\text{phys}} + O(a^p)$ for some $p > 0$, uniformly in a for a small enough. This is a statement about the approach to the continuum limit (the smoothness of the a -dependence), not about the existence of the limit itself.

What is NOT required. The Clay problem asks additionally for the construction of a continuum theory satisfying the Wightman axioms. From the lattice-first viewpoint, this is a separate mathematical question—interesting, but not physically necessary. The physics (the mass gap, the hadron spectrum, the confinement) is already contained in the lattice theory at finite a . The continuum limit is a mathematical idealisation: the statement that the rendering can be refined to infinite resolution. Whether or not the idealisation exists as a rigorous mathematical object does not change the physics at finite a .

Why the Clay formulation is physically misguided

The Clay problem conflates two separate questions:

Question 1 (physics): Does Yang–Mills theory confine and have a mass gap?

Question 2 (mathematics): Does the continuum limit of Yang–Mills theory exist as a rigorous QFT satisfying the Wightman axioms?

Question 1 is answered by lattice QCD: yes, to high numerical confidence, at every lattice spacing, with rendering-stable results. What remains is to upgrade the numerical evidence to a rigorous proof on the lattice — a difficult but well-defined mathematical problem.

Question 2 is a question about the continuum rendering: can the $a \rightarrow 0$ limit be taken rigorously, producing a mathematical object on $\mathbb{R}^4$ with exact Poincaré invariance? This is a question about whether the screen image has a mathematically well-defined infinite-resolution limit. The answer is important for pure mathematics (it would establish the existence of a non-trivial interacting QFT in four dimensions) but irrelevant for physics (the physics is at finite a , not at $a = 0$).

The Clay prize bundles Questions 1 and 2 together: the prize is awarded for proving both existence (Q2) and mass gap (Q1). From the lattice-first viewpoint, Question 1 should be the prize problem, and the correct formulation should be:

Prove that lattice Yang–Mills theory with gauge group $SU(N)$ on a hypercubic lattice in $d = 4$ has a mass gap $\Delta(a, L) > 0$ that survives the thermodynamic limit $L \rightarrow \infty$ and is rendering-stable: $\Delta(a) = \Delta_{\text{phys}} + O(a^2)$ as $a \rightarrow 0$.

This formulation:

(i) Does not require the continuum limit to exist as a Wightman QFT.

- (ii) Does not require exact Poincaré invariance (a rendered symmetry).
- (iii) Does require the physically meaningful content: a positive mass gap at every lattice spacing, surviving infinite volume, and stable under changes of a .
- (iv) Is closer to being proved: the lattice transfer matrix $T > 0$ is established, the finite-volume gap is proven by Perron–Frobenius, and the rendering-stability is confirmed numerically. What remains is the thermodynamic limit and the uniformity in a .

Current status of rigorous results

For completeness, we summarise the state of rigorous results on the lattice Yang–Mills mass gap:

In $d = 2$: the continuum Yang–Mills theory is exactly soluble (Migdal 1975, Witten 1991), and the mass gap can be computed analytically. The theory is “trivial” in the sense that it has no propagating degrees of freedom (no transverse gluons in $d = 2$), so the mass gap is of a different character than in $d = 4$.

In $d = 3$: partial results exist. The lattice theory is superrenormalisable (the coupling g^2 has dimensions of mass), and Balaban (1985–1989) constructed the continuum limit of the lattice theory in a finite volume, establishing UV stability. The mass gap in the thermodynamic limit is not yet proven.

In $d = 4$: the lattice theory is asymptotically free. Balaban’s block-spin renormalization group programme, which constructs the effective action at each scale by integrating out fluctuations shell by shell, is the most developed rigorous approach. The programme establishes that the lattice theory is well-defined (positive transfer matrix, reflection positivity) and that the coupling runs logarithmically (asymptotic freedom on the lattice). The mass gap and the thermodynamic limit remain open.

The gap between the numerical evidence (the mass gap is ~ 1710 MeV, computed to $\sim 5\%$ precision, rendering-stable across multiple lattice spacings) and the rigorous proof is primarily technical: the tools of constructive field theory (cluster expansions, polymer representations, renormalization group bounds) must be extended from the superrenormalisable setting ($d = 2, 3$) to the asymptotically free setting ($d = 4$).

The Fock-Born diagnosis. The difficulty of the Clay problem is not physical: the physics is clear (confinement exists; the mass gap is ~ 1710 MeV; the rendering is stable). The difficulty is mathematical: proving that the rendering-stable lattice gap survives the thermodynamic limit requires controlling the long-distance correlations of the gauge field, which fall off as $e^{-\Delta r}$ (massive) but involve $O(10^9)$ coupled degrees of freedom. The physical insight (confinement implies a mass gap) is clear; the mathematical implementation (a rigorous bound $\Delta > 0$ uniform in L) is hard.

The lattice-first viewpoint does not make the mathematics easier. But it clarifies what needs to be proved: a spectral gap of a lattice Hamiltonian, uniform in volume, stable under rescaling. Not a continuum limit. Not a Wightman QFT. Not exact Poincaré invariance. Just a gap, on a lattice, that doesn’t close.

42 General Relativity from the Action Principle

General relativity is not a separate theory bolted onto mechanics. It is the inevitable consequence of one idea: write the action of massive particles relative to a metric $g_{\mu\nu}(x)$, and allow $g_{\mu\nu}$ itself to vary in the extremisation. The Einstein equation is the saddle point of the combined action, and it is a self-consistency equation—not a causal loop.

42.1 Massive particles on flat spacetime

Consider N massive particles with masses m_A ($A = 1, \dots, N$), each tracing a worldline $x_A^\mu(\lambda_A)$ in flat spacetime with metric $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$. The proper time along worldline A :

$$d\tau_A = \sqrt{-\eta_{\mu\nu} dx_A^\mu dx_A^\nu} / c. \quad (42.1)$$

The action for N free particles:

$$S_{\text{free}}[\{x_A\}] = - \sum_{A=1}^N m_A c^2 \int d\tau_A = - \sum_{A=1}^N m_A c \int \sqrt{-\eta_{\mu\nu} \dot{x}_A^\mu \dot{x}_A^\nu} d\lambda_A. \quad (42.2)$$

Each particle contributes minus its rest energy times its proper time. The total action is the total mass-weighted proper time (with a sign so that extremising S *maximises* the proper time: the principle of maximal ageing).

Varying S_{free} with respect to $x_A^\mu(\lambda_A)$:

$$\frac{\delta S}{\delta x_A^\mu} = 0 \quad \implies \quad \frac{d^2 x_A^\mu}{d\tau_A^2} = 0 : \quad (42.3)$$

uniform motion in a straight line. Newton's first law, derived from the action principle.

This is the empty game: flat metric, straight worldlines, no gravity.

42.2 Einstein's happiest thought: let $g_{\mu\nu}$ vary

Einstein's "happiest thought" (1907) was the equivalence principle: a freely falling observer cannot distinguish gravity from the absence of gravity. In the language of the action: the observer's proper time $d\tau = \sqrt{-g_{\mu\nu} dx^\mu dx^\nu} / c$ should involve a metric $g_{\mu\nu}(x)$ that varies from point to point, not the fixed $\eta_{\mu\nu}$.

The promotion is immediate. Replace $\eta_{\mu\nu}$ with a general symmetric tensor field $g_{\mu\nu}(x)$:

$$S_{\text{matter}}[g, \{x_A\}] = - \sum_{A=1}^N m_A c \int \sqrt{-g_{\mu\nu}(x_A) \dot{x}_A^\mu \dot{x}_A^\nu} d\lambda_A. \quad (42.4)$$

At each event on worldline A , the proper time involves the metric $g_{\mu\nu}$ at that event. The action now depends on two sets of variables: the worldlines $\{x_A^\mu(\lambda_A)\}$ and the metric $g_{\mu\nu}(x)$.

Varying $g_{\mu\nu}$: the stress-energy tensor. The sensitivity of the matter action to changes in the metric defines the stress-energy tensor:

$$T_{\mu\nu}(x) = - \frac{2}{\sqrt{-g}} \frac{\delta S_{\text{matter}}}{\delta g^{\mu\nu}(x)}. \quad (42.5)$$

For a point particle: $T_{\mu\nu}(x) = \sum_A m_A c \int d\tau_A u_{A,\mu} u_{A,\nu} \delta^{(4)}(x - x_A(\tau_A)) / \sqrt{-g}$.

$T_{\mu\nu}$ is not an independent input. It is a functional of $g_{\mu\nu}$: it tells the metric which direction to bend in order to reduce the matter's total cost (increase the total proper time). Change the metric, and $T_{\mu\nu}$ changes.

Varying the worldlines: the geodesic equation. Extremising S_{matter} with respect to x_A^μ at fixed $g_{\mu\nu}$:

$$\frac{d^2 x_A^\mu}{d\tau_A^2} + \Gamma_{\rho\sigma}^\mu(x_A) \frac{dx_A^\rho}{d\tau_A} \frac{dx_A^\sigma}{d\tau_A} = 0, \quad (42.6)$$

where the Christoffel symbol $\Gamma_{\rho\sigma}^\mu = \frac{1}{2}g^{\mu\lambda}(\partial_\rho g_{\sigma\lambda} + \partial_\sigma g_{\rho\lambda} - \partial_\lambda g_{\rho\sigma})$ encodes the first derivatives of the metric. The geodesic equation: the worldline of maximal proper time in the given metric. When $g_{\mu\nu} = \eta_{\mu\nu}$: $\Gamma = 0$ and we recover the straight line (42.3).

But the matter action alone does not penalise curvature. Any $g_{\mu\nu}$ that makes the worldlines geodesic is equally good, no matter how wildly curved. The metric is undetermined—we need a cost for curvature.

42.3 Curvature as regularization: the Einstein–Hilbert action

The variational problem $\delta S_{\text{matter}}/\delta g^{\mu\nu} = 0$ is ill-posed: too many metrics extremise the matter action. This is the same situation as an under-regularised optimisation problem: without a penalty term, the solution is not unique.

We add a regularizer: a term in the action that penalises deviation from flatness. The regularizer must be a $\text{Diff}(M)$ -invariant scalar (to respect the coordinate freedom), built from $g_{\mu\nu}$ and its derivatives. The simplest such scalar with at most two derivatives is the Ricci scalar R :

$$R = g^{\mu\nu} R_{\mu\nu}, \quad R_{\mu\nu} = R^\rho_{\mu\rho\nu}, \quad R^\mu_{\nu\rho\sigma} = \partial_\rho \Gamma_{\nu\sigma}^\mu - \partial_\sigma \Gamma_{\nu\rho}^\mu + \Gamma_{\rho\lambda}^\mu \Gamma_{\nu\sigma}^\lambda - \Gamma_{\sigma\lambda}^\mu \Gamma_{\nu\rho}^\lambda. \quad (42.7)$$

The Riemann tensor $R^\mu_{\nu\rho\sigma}$ is the part of the second derivatives of $g_{\mu\nu}$ that cannot be removed by any coordinate transformation. The Ricci tensor $R_{\mu\nu}$ is its trace; the Ricci scalar R is the trace of the trace. Flat spacetime: $R = 0$. Curved spacetime: $R \neq 0$.

The Einstein–Hilbert action:

$$S_{\text{EH}}[g] = \frac{c^3}{16\pi G} \int d^4x \sqrt{-g} R. \quad (42.8)$$

Newton's constant G is the tuning parameter: it controls the strength of the regularization. Large $1/G$ (stiff spacetime): the curvature cost is high, and the metric stays close to flat. Small $1/G$: the curvature cost is low, and the metric bends easily. In our universe: $G \approx 6.67 \times 10^{-11} \text{ N} \cdot \text{m}^2/\text{kg}^2$ —very small, meaning spacetime is very stiff.

Higher-curvature terms (R^2 , $R_{\mu\nu}R^{\mu\nu}$, etc.) are higher-order regularizers. They are suppressed by additional powers of the Planck length $\ell_P = \sqrt{\hbar G/c^3} \sim 10^{-35} \text{ m}$ and are irrelevant operators in the sense of section ??: they contribute $o(a/\lambda)$ corrections to Born-layer observables and vanish in the stable rendering.

42.4 The combined saddle point: the Einstein equation

The total action:

$$S[g, \{x_A\}] = S_{\text{EH}}[g] + S_{\text{matter}}[g, \{x_A\}] = \frac{c^3}{16\pi G} \int d^4x \sqrt{-g} R - \sum_A m_A c^2 \int d\tau_A. \quad (42.9)$$

A competition between two costs: the curvature cost ($\int \sqrt{-g} R$: penalises deviation from flat) and the proper time reward ($\sum m_A c^2 \int d\tau_A$: rewards ageing).

Varying S with respect to $g^{\mu\nu}$:

$$\frac{\delta S_{\text{EH}}}{\delta g^{\mu\nu}} + \frac{\delta S_{\text{matter}}}{\delta g^{\mu\nu}} = 0, \quad (42.10)$$

which gives:

$$\boxed{G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \frac{8\pi G}{c^4} T_{\mu\nu}.} \quad (42.11)$$

The Einstein equation. The left side is the Einstein tensor $G_{\mu\nu}$: a geometric object built from the Riemann tensor. The right side is $T_{\mu\nu}$: the stress-energy tensor, defined by (42.5).

This is a self-consistency equation. Both sides depend on $g_{\mu\nu}$. The left side $G_{\mu\nu}[g]$ depends on g through the Riemann tensor and its contractions. The right side $T_{\mu\nu}[g, \psi]$ depends on g through the matter coupling: $g^{\mu\nu}$ contracts indices, $\Gamma_{\rho\sigma}^\mu$ enters covariant derivatives, $\sqrt{-g}$ provides the volume element.

The equation does not say “given $T_{\mu\nu}$, compute $g_{\mu\nu}$.” It says: find the $g_{\mu\nu}$ such that $G_{\mu\nu}[g]$ and $T_{\mu\nu}[g, \psi]$ are consistent. Both sides adjust together. Neither is prior. This is an equation of state—a self-consistency condition—not a causal mechanism.

The Bianchi identity and the redundancy. The Einstein tensor satisfies the contracted Bianchi identity: $\nabla_\mu G^{\mu\nu} \equiv 0$ (identically, for any $g_{\mu\nu}$). Combined with the Einstein equation: $\nabla_\mu T^{\mu\nu} = 0$, which is the conservation of energy-momentum. But this conservation law also follows from varying S_{matter} with respect to diffeomorphisms—it is the Noether current of coordinate invariance. The geodesic equation (42.6) follows from $\nabla_\mu T^{\mu\nu} = 0$ for point particles. The two “instructions” in Wheeler’s dictum (section 42.6) are not independent: the Bianchi identity connects them.

42.5 Einstein’s uneasiness dissolved

Einstein described his field equations as architecturally asymmetric: the left side ($G_{\mu\nu}$, built from pure geometry) is a “palace of marble”; the right side ($T_{\mu\nu}$, describing matter) is a “hovel of wood.” He spent three decades trying to derive the matter from the geometry.

From the action principle, the uneasiness is unfounded:

$$G_{\mu\nu} = -\frac{16\pi G}{c^3\sqrt{-g}} \frac{\delta S_{\text{EH}}}{\delta g^{\mu\nu}} \quad (\text{derivative of the regularizer}), \quad (42.12)$$

$$T_{\mu\nu} = -\frac{2}{\sqrt{-g}} \frac{\delta S_{\text{matter}}}{\delta g^{\mu\nu}} \quad (\text{derivative of the objective}). \quad (42.13)$$

Both sides are functional derivatives of the *same* action $S[g, \psi]$ with respect to the *same* variable $g^{\mu\nu}$, set equal by the *same* extremum condition $\delta S = 0$. The “marble” is the gradient of the regularizer; the “straw” is the gradient of the objective. Both are computed from the action; both are needed for the extremum; neither is more fundamental than the other.

$T_{\mu\nu}$ is not a phenomenological, ugly, put-in-by-hand description of “what matter is.” It is a principled description of **what matter wants from the metric**: the direction in which the metric should bend to increase the total proper time. The regularizer ($G_{\mu\nu}$) resists; the objective ($T_{\mu\nu}$) demands. The Einstein equation is the balance.

There was never a palace and a hovel. There was one action, one variational principle, one building.

42.6 Wheeler’s dictum, corrected

John Wheeler’s famous summary:

“Spacetime tells matter how to move;
matter tells spacetime how to curve.”

is wrong in three instructive ways.

1. *There is one action, not two agents.* Wheeler’s phrasing implies two mechanisms acting in sequence: matter curves spacetime (Einstein equation), then spacetime guides matter (geodesic equation). But $\delta S = 0$ is one equation with two projections. The Bianchi identity $\nabla_\mu G^{\mu\nu} \equiv 0$ means that the Einstein equation alone already implies geodesic motion for the matter. Wheeler’s two instructions are one equation.

2. *The metric is not an agent; it is a variable.* “Spacetime tells matter how to move” attributes agency to $g_{\mu\nu}$. But $g_{\mu\nu}$ is a symmetric tensor field determined by the extremum of $S[g, \psi]$. It does not “tell” anything; it is the solution of a self-consistency equation. (In the Fock-Born framework: $g_{\mu\nu}$ is a Born-layer output, not an agent.)

3. *There are no causal arrows.* The standard reading assigns arrows: matter $\rightarrow$ curvature $\rightarrow$ motion. The action formulation has no arrows: g and $\{x_A\}$ are determined simultaneously by the joint extremum. Asking which causes which is like asking whether $\partial f / \partial x = 0$ causes $\partial f / \partial y = 0$ in the extremum of $f(x, y)$.

The corrected dictum:

Mass curves spacetime at the cheapest possible curvature cost,
so that mass can age as fast as possible.

One sentence, one agent (the joint system $g + \{x_A\}$), one optimisation:

Phrase	Referent in $S[g, \{x_A\}]$
“Mass curves spacetime”	$T_{\mu\nu} \neq 0$ sources $G_{\mu\nu} = (8\pi G/c^4)T_{\mu\nu}$
“cheapest possible curvature cost”	$S_{\text{EH}} = \frac{c^3}{16\pi G} \int \sqrt{-g} R$; stiffness $\propto 1/G$
“age as fast as possible”	Maximise $\sum_A m_A c^2 \int d\tau_A$; the geodesic is the path of maximal proper time
“one optimisation”	$\delta S[g, \{x_A\}] = 0$ jointly in $g^{\mu\nu}$ and x_A^μ

42.7 Black holes: the Schwarzschild geometry completed

The Schwarzschild metric (42.21) has a coordinate singularity at $r = r_s$ (the “Schwarzschild radius” or “event horizon”) and a genuine curvature singularity at $r = 0$. The region $r < r_s$ is the interior of a black hole: a region of spacetime from which no causal signal can escape to infinity.

The horizon is a Born-layer surface. The event horizon $r = r_s$ is not a physical barrier. An infalling observer crosses it without experiencing anything locally extraordinary: the tidal forces at $r = r_s$ are $\sim GM/(c^2 r_s^3) = c^6/(4G^2 M^2)$, which for a solar-mass black hole is modest ($\sim 10^{10} \text{ m}^{-2}$) and for a supermassive black hole ($M \sim 10^9 M_\odot$) is negligible. The horizon is defined globally: it is the boundary of the causal past of future null infinity. No local measurement can detect it.

In the Fock-Born framework: the horizon is a property of the Born-layer rendering $g_{\mu\nu}$, not a local feature of the Fock-layer engine. The lattice Hamiltonian $H(a)$ does not “know” about the horizon; the horizon emerges in the macroscopic rendering of $g_{\mu\nu}$ as the surface where the timelike Killing vector $\partial/\partial t$ becomes null.

The Penrose diagram. The maximal analytic extension of the Schwarzschild geometry

(Kruskal–Szekeres coordinates) reveals four regions: the exterior (Region I, where we live), the black hole interior (Region II, the future singularity), a white hole (Region III, the past singularity), and a second exterior (Region IV). The Penrose diagram compresses the entire causal structure into a finite square, with null rays at 45° .

The physical black hole formed by gravitational collapse has only Regions I and II: the white hole and the second exterior are artefacts of the maximal extension and do not form in a realistic collapse.

The Kerr metric. Astrophysical black holes rotate. The Kerr metric (1963) describes a rotating black hole with mass M and angular momentum $J = aGM/c$:

$$ds^2 = -\left(1 - \frac{r_s r}{\Sigma}\right) c^2 dt^2 - \frac{2r_s r a \sin^2 \theta}{\Sigma} c dt d\varphi + \frac{\Sigma}{\Delta} dr^2 + \Sigma d\theta^2 + \frac{A \sin^2 \theta}{\Sigma} d\varphi^2, \quad (42.14)$$

where $\Sigma = r^2 + a^2 \cos^2 \theta$, $\Delta = r^2 - r_s r + a^2$, and $A = (r^2 + a^2)^2 - a^2 \Delta \sin^2 \theta$. The outer horizon is at $r_+ = (r_s + \sqrt{r_s^2 - 4a^2})/2$. The bound $a \leq r_s/2$ (i.e., $J \leq GM^2/c$) is the Kerr bound: a black hole with $a > r_s/2$ would have a naked singularity, which the cosmic censorship conjecture forbids.

The no-hair theorem (Israel, Carter, Robinson): a stationary black hole in GR is uniquely characterised by three parameters — mass M , angular momentum J , and electric charge Q — regardless of the details of the matter that formed it. All other information about the collapsing matter is radiated away as gravitational waves during the collapse. The three parameters M , J , Q are Born-layer observables: what a distant observer measures from the gravitational field, the frame-dragging effect, and the electric field.

Observational evidence. Black holes are observed through their Born-layer effects:

Stellar-mass black holes ($M \sim 5\text{--}100 M_\odot$): detected in X-ray binaries (accretion disk radiation) and gravitational wave mergers (LIGO/Virgo). GW150914 confirmed the merger of two black holes and the ringdown to a Kerr black hole, matching the no-hair prediction.

Supermassive black holes ($M \sim 10^6\text{--}10^{10} M_\odot$): detected at galactic centres through stellar orbits (the 2020 Nobel Prize to Genzel and Ghez for observing stellar orbits around Sgr A*) and direct imaging (the Event Horizon Telescope’s 2019 image of M87*, showing the photon ring and shadow).

The EHT image: the dark shadow in the EHT image of M87* is the photon capture cross section of the black hole—the set of photon trajectories that do not escape to infinity. The shadow radius $r_{\text{shadow}} \approx 5.2 GM/c^2$ is a Born-layer observable computed from null geodesics in the Kerr metric. The agreement between the observed shadow size and the GR prediction confirms the Kerr geometry in the strong-field regime.

42.8 Dark energy and the accelerating expansion

The most general action consistent with diffeomorphism invariance and the equivalence principle includes a cosmological constant term:

$$S = \frac{1}{16\pi G} \int d^4x \sqrt{-g} (R - 2\Lambda) + S_{\text{matter}}. \quad (42.15)$$

The cosmological constant Λ appears as a constant contribution to the energy density of spacetime: a uniform energy density $\rho_\Lambda = \Lambda c^2/(8\pi G)$ and a negative pressure $p_\Lambda = -\rho_\Lambda c^2$ (the equation of state $w = p/(\rho c^2) = -1$).

The Friedmann equation. For a homogeneous, isotropic universe with scale factor $a(t)$ (the Friedmann–Lemaître–Robertson–Walker metric $ds^2 = -c^2 dt^2 + a^2(t)[dr^2/(1 - kr^2) + r^2 d\Omega^2]$), the Einstein equation reduces to:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}, \quad (42.16)$$

where ρ is the total energy density (matter + radiation) and $k = 0, \pm 1$ is the spatial curvature. The Hubble parameter $H(t) = \dot{a}/a$ measures the expansion rate.

The equation of state and $1 + w$. A perfect fluid with energy density ρ and pressure p has equation of state $w = p/(\rho c^2)$. The value of w determines how the fluid responds to the cosmic expansion. The continuity equation $\dot{\rho} + 3H(\rho + p/c^2) = 0$ (the $\nabla_\mu T^{\mu 0} = 0$ component of stress-energy conservation) gives:

$$\dot{\rho} + 3H\rho(1 + w) = 0 \quad \implies \quad \rho \propto a^{-3(1+w)}. \quad (42.17)$$

The quantity $1 + w$ controls everything: it is the exponent that determines how the energy density dilutes as the universe expands.

Component	w	$1+w$	$\rho \propto$	Physical meaning
Dust (matter)	0	1	a^{-3}	Number density dilutes as volume a^3 grows
Radiation	$\frac{1}{3}$	$\frac{4}{3}$	a^{-4}	Number density $\times$ redshift: each photon also loses energy $\propto 1/a$
Λ (cosm. const.)	-1	0	a^0	Energy density is constant: new volume is created with the same ρ_Λ
Curvature	$-\frac{1}{3}$	$\frac{2}{3}$	a^{-2}	The $-kc^2/a^2$ term in Friedmann
Phantom	< -1	< 0	grows	Energy density <i>increases</i> with expansion; violates the null energy condition

The three physically important thresholds are:

$1 + w = 0$ (*the cosmological constant, $w = -1$*). The energy density does not dilute. As the universe expands and new comoving volume is created, it arrives pre-loaded with the same energy density ρ_Λ . This is because the cosmological constant is a property of the action $\int \sqrt{-g} \Lambda$ —it scales with the volume element $\sqrt{-g} d^4x$, not with the matter content. In the Fock-Born framework: ρ_Λ is a parameter of the engine, not a diluting substance. It does not redshift, scatter, or cluster. It is the same everywhere and at all times.

$1 + 3w = 0$ ($w = -1/3$): *the acceleration threshold*. The second Friedmann equation (the acceleration equation):

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\rho(1 + 3w). \quad (42.18)$$

Acceleration ($\ddot{a} > 0$) requires $1 + 3w < 0$, i.e., $w < -1/3$. Matter ($w = 0$) and radiation ($w = 1/3$) both decelerate the expansion. Only a component with sufficiently negative pressure ($w < -1/3$) can accelerate it. The cosmological constant ($w = -1$) satisfies this comfortably.

$1 + w < 0$ ($w < -1$): *the phantom divide*. If $1 + w < 0$, the energy density *increases* as the universe expands—(42.17) gives $\rho \propto a^{|3(1+w)|}$ growing with a . This violates the null energy condition $\rho + p/c^2 \geq 0$ (equivalently $T_{\mu\nu}k^\mu k^\nu \geq 0$ for all null k^μ), which is a sufficient condition for the focusing of null geodesics (the Penrose singularity theorem) and for the positivity of the Bondi mass. A component with $w < -1$ (“phantom dark energy”) is pathological: the vacuum

energy grows without bound, the universe ends in a “Big Rip” (all bound structures are torn apart at finite time), and the theory has ghost-like instabilities at the quantum level.

Current observations constrain $w = -1.00 \pm 0.05$: no evidence for $1 + w \neq 0$. The cosmological constant ($1 + w = 0$ exactly) remains the simplest explanation.

The observational evidence. In 1998, Riess et al. and Perlmutter et al. discovered that the expansion of the universe is accelerating: distant Type Ia supernovae are fainter than expected in a decelerating universe, indicating that the expansion rate $\dot{a}/a$ is increasing with time. The acceleration requires a component with negative pressure — “dark energy” — contributing $\sim 68\%$ of the total energy density.

The combined evidence from:

Type Ia supernovae: standardisable candles at cosmological distances. The luminosity distance $d_L(z)$ as a function of redshift z deviates from the matter-dominated prediction at $z \gtrsim 0.3$, indicating acceleration.

The cosmic microwave background: the angular power spectrum (Planck satellite) constrains the total energy density $\Omega_{\text{tot}} = 1.000 \pm 0.002$ and the matter fraction $\Omega_m = 0.315 \pm 0.007$, requiring $\Omega_\Lambda = 1 - \Omega_m \approx 0.685$.

Baryon acoustic oscillations: the characteristic scale of the BAO feature in the galaxy correlation function provides a standard ruler at multiple redshifts, independently confirming the Λ CDM model.

All three datasets are consistent with a cosmological constant Λ with $w = -1.00 \pm 0.05$: no evidence for time variation.

The cosmological constant problem. The observed value:

$$\rho_\Lambda^{\text{obs}} \sim (2.3 \times 10^{-3} \text{ eV})^4 \sim 10^{-47} \text{ GeV}^4. \quad (42.19)$$

The quantum field theory prediction for the vacuum energy density (the $\mathbf{A}_0$ coefficient of the heat kernel, section 39):

$$\rho_\Lambda^{\text{QFT}} \sim \frac{1}{16\pi^2} \sum_i \eta_i \frac{1}{a^4} \sim \frac{1}{a^4} \sim M_P^4 \sim 10^{76} \text{ GeV}^4 \quad (\text{if } a \sim \ell_P). \quad (42.20)$$

The ratio $\rho_\Lambda^{\text{obs}}/\rho_\Lambda^{\text{QFT}} \sim 10^{-123}$: the most extreme discrepancy in physics.

In the lattice-first viewpoint: the bare vacuum energy $E_0(a)$ is a parameter of the engine, adjusted by the rendering-stability condition to produce $\rho_\Lambda^{\text{phys}} = \rho_\Lambda^{\text{obs}}$. The adjustment requires a cancellation of ~ 123 decimal places between $E_0(a)$ and the quantum corrections. This is the cosmological constant problem: not an inconsistency (the condition can always be satisfied) but a statement about the engine’s calibration. The function $E_0(a)$ must be extraordinarily fine-tuned.

The Fock-Born reading. The cosmological constant is a Born-layer observable: it determines the acceleration of the cosmic expansion, measured by the supernova Hubble diagram, the CMB power spectrum, and the BAO scale. The Fock-layer source is the bare vacuum energy $E_0(a)$ plus the quantum corrections from all matter fields ($\mathbf{A}_0$). The fine-tuning required to match the observed value is a property of the engine’s calibration — the most extreme instance of the rendering-stability condition.

Whether this fine-tuning has a dynamical explanation (a relaxation mechanism, an anthropic selection, a symmetry protecting the cosmological constant) or is simply a brute fact about the

engine is the deepest open question in fundamental physics. The Fock-Born framework does not resolve it; it states it precisely: the function $E_0(a)$ must be whatever it needs to be to produce $\rho_\Lambda^{\text{obs}}$ as a Born-layer output.

42.9 The rendering postulate and the Fock-Born reading

In the Fock-Born framework, $g_{\mu\nu}$ is a Born-layer output. It is the macroscopic, classical, highly degenerate rendering of the graviton field—a composite observable with degeneracy $\sim e^{S_{\text{BH}}/k_B}$ (where S_{BH} is the Bekenstein–Hawking entropy), pushed to the extreme classical end of the Born hierarchy.

The *rendering postulate*: the observer experiences the eigenvalue $g_{\mu\nu}(x)$ as the local geometry of spacetime. Proper time: $d\tau = \sqrt{-g_{\mu\nu}dx^\mu dx^\nu}/c$. Proper length: $dl = \sqrt{g_{ij}dx^i dx^j}$. These are Born-layer numbers: what clocks and rods report.

The rendering postulate is the bridge between the Fock layer (where $g_{\mu\nu}$ is computed as the classical limit of a quantum field) and the Born layer (where the observer experiences it as geometry). The classical Einstein equation is the saddle-point approximation to the quantum theory—the tree-level effective equation for $\langle \hat{g}_{\mu\nu} \rangle$.

The one-way valve. The metric $g_{\mu\nu}$ is a Born-layer output. It cannot be recycled as a Fock-layer input. Attempting to quantise gravity by writing a Hamiltonian H on a Hilbert space with a dynamical $g_{\mu\nu}$ is recycling the rendering back into the engine. This is the deepest violation of the Fock-Born architecture, and it is the origin of the conceptual difficulties of quantum gravity (the problem of time, the problem of observables, the non-renormalisability of the Einstein–Hilbert action at two loops).

The Fock layer operates on a flat lattice with spacing a . The metric $g_{\mu\nu}$ emerges in the Born layer at $\lambda \gg a$, as a macroscopic rendering of the graviton field’s vacuum expectation value. The lattice does not “live on” $g_{\mu\nu}$; it lives on the flat computational substrate. $g_{\mu\nu}$ is what the screen shows, not what the engine runs on.

Numerical relativity confirms this. Every numerical relativity code (solving the Einstein equation for binary black hole mergers, neutron star collisions, gravitational wave templates) operates on a flat computational grid with spacing Δx . The code evolves $g_{\mu\nu}(\mathbf{x}, t)$ as data on this flat grid. The grid does not curve; the data on the grid encodes curvature. The grid is the Fock layer; $g_{\mu\nu}$ is the Born-layer output. Every LIGO template is a Born-layer rendering computed on a flat Fock-layer grid.

42.10 The Schwarzschild solution: a Born-layer output

The simplest non-trivial solution of the Einstein equation (42.11) is the Schwarzschild metric: the unique spherically symmetric, static, vacuum solution ($T_{\mu\nu} = 0$ outside the source).

The metric. In Schwarzschild coordinates (t, r, θ, φ) :

$$ds^2 = -\left(1 - \frac{r_s}{r}\right)c^2 dt^2 + \left(1 - \frac{r_s}{r}\right)^{-1} dr^2 + r^2 d\Omega^2, \quad (42.21)$$

where $r_s = 2GM/c^2$ is the Schwarzschild radius and $d\Omega^2 = d\theta^2 + \sin^2 \theta d\varphi^2$ is the metric on the unit 2-sphere. At $r \gg r_s$: $g_{\mu\nu} \rightarrow \eta_{\mu\nu}$ and the metric is flat. At $r = r_s$: the coordinate t diverges but the geometry is regular (the “singularity” is a coordinate artefact, removed by switching to Eddington–Finkelstein or Kruskal–Szekeres coordinates). At $r = 0$: a genuine curvature singularity where $R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} \rightarrow \infty$.

The Newtonian limit. At $r \gg r_s$, the metric component $g_{00} \approx -(1 - 2GM/(c^2 r))$ recovers the Newtonian gravitational potential $\Phi = -GM/r$:

$$g_{00} \approx -(1 + 2\Phi/c^2), \quad \Phi = -\frac{GM}{r}. \quad (42.22)$$

The geodesic equation (42.6) in this limit reduces to Newton's second law $\ddot{\mathbf{x}} = -\nabla\Phi = -GM\hat{r}/r^2$: gravity as a force is the Born-layer rendering of geodesic motion in the Schwarzschild metric.

Born-layer predictions. The Schwarzschild metric is a Born-layer output: the macroscopic, classical rendering of the graviton field around a spherical mass. It predicts Born-layer observables that have been measured to high precision:

Gravitational time dilation. A clock at radial coordinate r ticks at rate $d\tau = \sqrt{-g_{00}} dt = \sqrt{1 - r_s/r} dt$ relative to a clock at infinity. Clocks closer to the mass run slower. This is measured by GPS satellites: the onboard atomic clocks at altitude $\sim 20,200$ km run $\sim 45 \mu\text{s/day}$ faster than ground clocks due to the weaker gravitational field, compensated by $\sim 7 \mu\text{s/day}$ slower due to their orbital speed (special relativistic time dilation). The net correction of $\sim 38 \mu\text{s/day}$ is a Born-layer observable: what GPS receivers report when the correction is applied. Without it, GPS positions would drift by ~ 10 km/day.

Gravitational redshift. A photon climbing out of a gravitational well loses energy: $\nu_{\text{received}}/\nu_{\text{emitted}} = \sqrt{g_{00}(r_{\text{emit}})/g_{00}(r_{\text{recv}})}$. Measured by Pound–Rebka (1959): a γ -ray climbing 22.5 m in the Harvard tower is redshifted by $\Delta\nu/\nu = gh/c^2 \approx 2.5 \times 10^{-15}$, in agreement with GR to 1%.

Perihelion precession. The geodesic equation in the Schwarzschild metric gives a precessing ellipse. For Mercury: 43 arcseconds per century beyond the Newtonian prediction, confirming GR to 0.1%.

Gravitational lensing. A photon passing a mass at impact parameter b is deflected by angle $\delta\theta = 4GM/(c^2 b)$: twice the Newtonian prediction. Confirmed by Eddington (1919) during a solar eclipse, and now routinely used in gravitational lensing surveys to map the dark matter distribution.

Each of these is a Born-layer observable: a number extracted from the rendering of $g_{\mu\nu}$ by a macroscopic detector (a clock, a spectrometer, a telescope). The Fock-layer source is the Schwarzschild metric; the Born-layer output is what the instrument reports.

42.11 Gravitational waves: ripples in the Born-layer output

Linearising the Einstein equation around flat spacetime ($g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}/M_P$, with $|h_{\mu\nu}| \ll 1$) gives the wave equation for the metric perturbation:

$$\square \bar{h}_{\mu\nu} = -\frac{16\pi G}{c^4} T_{\mu\nu}, \quad (42.23)$$

where $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$ is the trace-reversed perturbation and $\square = -\partial_t^2/c^2 + \nabla^2$ is the d'Alembertian. In vacuum ($T_{\mu\nu} = 0$): this is a wave equation with solutions propagating at the speed of light.

The two polarisations. In transverse-traceless (TT) gauge, the gravitational wave has two independent polarisations, h_+ and $h_\times$, corresponding to the two helicity states $\lambda = \pm 2$ of a massless spin-2 field. A passing gravitational wave stretches space in one transverse direction while compressing it in the perpendicular direction. The strain $h = \Delta L/L$ is the fractional change in the distance between two free masses.

The quadrupole formula. A non-spherically-symmetric mass distribution with time-varying quadrupole moment $I_{ij}(t)$ radiates gravitational waves with strain at distance r :

$$h_{ij}^{\text{TT}}(t, r) = \frac{2G}{c^4 r} \ddot{I}_{ij}^{\text{TT}}(t_{\text{ret}}), \quad I_{ij} = \int d^3\mathbf{x} \rho(\mathbf{x}, t) \left(x_i x_j - \frac{1}{3} \delta_{ij} |\mathbf{x}|^2 \right), \quad (42.24)$$

where $t_{\text{ret}} = t - r/c$ is the retarded time. The emitted power:

$$P_{\text{GW}} = \frac{G}{5c^5} \langle \ddot{I}_{ij} \ddot{I}^{ij} \rangle. \quad (42.25)$$

The factor $G/c^5 \approx 2.6 \times 10^{-53} \text{ W}^{-1} \text{ s}^2$ is extraordinarily small: gravitational radiation is extremely weak. Only systems with enormous $\ddot{I}_{ij}$ (binary neutron stars, merging black holes, asymmetric supernova explosions) produce detectable gravitational waves.

The Hulse–Taylor binary: indirect detection. The binary pulsar PSR B1913+16, discovered by Hulse and Taylor (1974), loses orbital energy by emitting gravitational waves. The predicted orbital decay rate from (42.25): $\dot{P}_{\text{orb}}/P_{\text{orb}} = -(2.4 \times 10^{-12}) \text{ s}^{-1}$. The measured value agrees to better than 0.2%: a confirmation of gravitational radiation as a Born-layer observable (the cumulative shift in the pulsar’s arrival times).

LIGO: direct detection. The Laser Interferometer Gravitational-Wave Observatory detected the first gravitational wave signal on 14 September 2015: GW150914, from the merger of two black holes ($36 + 29 \rightarrow 62 M_{\odot}$) at a distance of ~ 1.3 billion light-years. The strain at Earth: $h \sim 10^{-21}$ —a displacement of $\Delta L \sim 10^{-18} \text{ m}$ over the 4 km arm length.

The detection is a Born-layer event: the interferometer submits a macroscopic observable (the difference in photon travel times between two arms), the Born rule renders an eigenvalue (the interference fringe shift), and the pattern of eigenvalues over time is matched to a template computed from the quadrupole formula (42.24) with full numerical-relativity corrections. The template is a Fock-layer computation (solving the Einstein equation on a flat computational grid). The detection is a Born-layer event (the photodetector click rate). The match between them confirms the Einstein equation to the strong-field, dynamical regime.

The Fock-Born diagnostic.

Quantity	Layer	What it is
Schwarzschild metric $g_{\mu\nu}$	Born	Macroscopic rendering of the graviton field
Metric perturbation $h_{\mu\nu}$	Born	Gravitational wave strain
Graviton ($h = \pm 2$)	Fock	Massless spin-2 excitation of $\hat{h}_{\mu\nu}$
Quadrupole formula	Fock	Computation of $\ddot{I}_{ij}^{\text{TT}}$
LIGO fringe shift	Born	Detector click pattern
Numerical relativity template	Fock	$g_{\mu\nu}$ solved on a flat grid
Perihelion precession	Born	Telescope measurement over decades
GPS time correction	Born	What the satellite reports

Every gravitational prediction follows the same pipeline as particle physics: the Fock layer computes a metric (by solving the Einstein equation on a flat lattice), and the Born layer renders the observable (the clock rate, the photon path, the interferometer fringe). The machinery is heavier—tensor equations instead of scalar ones—but the architecture is the same.

42.12 Coupling gravity to the Standard Model

The full action of particle physics plus gravity:

$$S[g, \Phi] = S_{\text{EH}}[g] + S_{\text{SM}}[g, \Phi], \quad (42.26)$$

where Φ denotes all Standard Model fields and the SM action is written on a general metric $g_{\mu\nu}$ by the minimal coupling prescription: $\eta_{\mu\nu} \rightarrow g_{\mu\nu}$, $\partial_\mu \rightarrow \nabla_\mu$, $d^4x \rightarrow d^4x \sqrt{-g}$.

The Einstein equation becomes:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}^{\text{SM}}, \quad (42.27)$$

where $T_{\mu\nu}^{\text{SM}} = -(2/\sqrt{-g}) \delta S_{\text{SM}} / \delta g^{\mu\nu}$ is the stress-energy tensor of all Standard Model fields. The matter side includes contributions from quarks, leptons, gauge bosons, and the Higgs field.

The hierarchy of couplings. The gravitational coupling G is extraordinarily weak compared to the other forces: $G_F^{-1/2} \sim 300 \text{ GeV}$ (weak), $\Lambda_{\text{QCD}} \sim 200 \text{ MeV}$ (strong), but $M_P = (8\pi G)^{-1/2} \sim 2.4 \times 10^{18} \text{ GeV}$ (gravity). The ratio $M_P/M_W \sim 10^{17}$ is the hierarchy problem stated as a ratio of scales: why is gravity 10^{34} times weaker than the weak force?

In the lattice-first viewpoint: the hierarchy is a property of the bare parameters $\theta(a)$. The Hamiltonian $H_{\text{SM+grav}}(a)$ contains the gravitational coupling $G(a)$ alongside the gauge couplings $g_s(a)$, $g(a)$, $g'(a)$. The rendering-stability condition determines how each coupling runs with a , but it does not explain their relative magnitudes at any particular a . The hierarchy is a boundary condition on the engine, not a consequence of its architecture.

The cosmological constant. The most general action consistent with diffeomorphism invariance includes a cosmological constant term:

$$S_\Lambda = -\frac{\Lambda}{8\pi G} \int d^4x \sqrt{-g}, \quad (42.28)$$

where Λ has dimensions of $(\text{length})^{-2}$. The observed value $\Lambda_{\text{obs}} \sim (10^{-3} \text{ eV})^4 / (M_P^2 c^4) \sim 10^{-52} \text{ m}^{-2}$ is $\sim 10^{120}$ times smaller than the natural scale $1/a^2$ (if $a \sim \ell_P$). This is the cosmological constant problem: the most extreme fine-tuning in physics, discussed in section 39.

In the Fock-Born framework: Λ is a bare parameter of the engine, tuned by the rendering-stability condition. The tuning is extreme but not inconsistent: the condition $\mathcal{O}_{\text{Born}} = \mathcal{O}_{\text{phys}} + o(a/\lambda)$ does not care whether $\theta(a)$ is “natural” or “unnatural.” It simply requires that the functions $\theta(a)$ produce the observed Born-layer output.

42.13 General relativity as a rendered symmetry

In the lattice-first viewpoint (section ??), diffeomorphism invariance—the gauge symmetry of general relativity—is a rendered symmetry: it holds at $\lambda \gg a$ because the diffeomorphism-breaking corrections are irrelevant operators suppressed by powers of a/λ .

The lattice has a fixed, flat, discrete structure: it breaks diffeomorphism invariance completely. The Hamiltonian $H(a)$ is constructed on this lattice with nearest-neighbour couplings. The Born-layer rendering—the metric $g_{\mu\nu}$ at wavelengths $\lambda \gg a$ —is diffeomorphism-invariant to accuracy $o(a/\lambda)^2$, because the leading corrections have mass dimension 6 (four derivatives of $g_{\mu\nu}$ contracted into a scalar):

$$\delta \mathcal{L} \sim a^2 (c_1 R^2 + c_2 R_{\mu\nu} R^{\mu\nu} + c_3 R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma}). \quad (42.29)$$

These are the same higher-curvature terms as in the effective field theory of gravity (section ??), now understood as lattice artefacts: they encode the breaking of diffeomorphism invariance by the substrate and are suppressed by $(a/\lambda)^2 \sim (E/M_P)^2 \sim 10^{-30}$ at LHC energies.

This completes the list of rendered symmetries:

Symmetry	Status on lattice	Rendered accuracy	Leading correction
Lorentz	Broken to cubic group	$o(a/\lambda)^2$	$a^2 k^4$ in dispersion
Diffeomorphism	Broken completely	$o(a/\lambda)^2$	$a^2 R^2$ terms
$SU(3)_c \times SU(2)_L \times U(1)_Y$	Exact (Wilson's lattice)	Exact	None
Translation	Broken to $a\mathbb{Z}^d$	$o(a/\lambda)^2$	Umklapp scattering

Gauge invariance is exact on the lattice (Wilson's construction maintains it at every a). Lorentz invariance, diffeomorphism invariance, and continuous translation invariance are all rendered: they emerge at $\lambda \gg a$ and are broken by $(a/\lambda)^2$ corrections that are too small to measure at current energies.

The universe, in this reading, is a lattice simulation. The lattice is the Fock layer. The laws of physics—Lorentz invariance, the Einstein equation, gauge invariance—are properties of the Born-layer rendering, not of the substrate. The substrate is simpler: a flat, discrete, finite array of coupled quantum degrees of freedom, with a Hamiltonian that is a polynomial of creation and annihilation operators. Everything else is rendered.

43 Quantum Gravity

This section proceeds in two stages. First, the semiclassical treatment: quantise the matter fields $\hat{\psi}$ on a fixed classical metric $g_{\mu\nu}$. This yields the Unruh effect, the Hawking effect, the Bekenstein–Hawking entropy, and the trace anomaly—deep results that follow from the action principle and the heat kernel without quantising gravity itself. Second, the quantum treatment of $g_{\mu\nu}$: we review existing approaches (the graviton, effective field theory, canonical quantisation, string theory) and examine each from the Fock-Born framework.

Part I: Semiclassical — quantum fields on classical $g_{\mu\nu}$

43.1 The semiclassical framework

	Level	$g_{\mu\nu}$	ψ
0	Full quantum gravity	Quantum: operator $\hat{g}$ on $\mathcal{F}$	Quantum: operator $\hat{\psi}$ on $\mathcal{F}$
1	Semiclassical	Classical: c-number, satisfies Einstein equations	Quantum: operator on Fock space built on g
2	Classical GR	Classical	Classical

Level 1 captures particle creation by gravitational fields, the Unruh and Hawking effects, the Bekenstein–Hawking entropy, and the trace anomaly—without quantising gravity itself.

The semiclassical action: $S[g, \psi] = S_{\text{EH}}[g] + S_{\text{matter}}[g, \psi]$. The semiclassical Einstein equation:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} \langle \hat{T}_{\mu\nu} \rangle_{\text{ren}}, \quad (43.1)$$

where $\langle \hat{T}_{\mu\nu} \rangle_{\text{ren}}$ is the renormalised expectation value of the quantum stress-energy in the state $|\Psi\rangle$. The metric is classical; the source is quantum.

43.2 Quantising the scalar field on $g_{\mu\nu}$

A scalar field on a curved background satisfies the Klein–Gordon equation $(\nabla_\mu \nabla^\mu - m^2 - \xi R)\phi = 0$, with action $S = -\frac{1}{2} \int d^4x \sqrt{-g} [g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi + (m^2 + \xi R)\phi^2]$.

The quantisation follows the flat-space procedure: expand $\hat{\phi}(x) = \sum_i [f_i(x) \hat{a}_i + f_i^*(x) \hat{a}_i^\dagger]$, where $\{f_i\}$ is a complete set of positive-frequency solutions. But on a general curved spacetime, the split into “positive” and “negative” frequency is not unique: there is no preferred time coordinate, hence no preferred notion of “a particle.”

The vacuum $|0_f\rangle$ defined by $\hat{a}_i |0_f\rangle = 0$ (using modes $\{f_i\}$) is a different state from the vacuum $|0_g\rangle$ defined by $\hat{b}_i |0_g\rangle = 0$ (using a different set of modes $\{g_i\}$). The two sets are related by a Bogoliubov transformation:

$$g_i = \sum_j (\alpha_{ij} f_j + \beta_{ij} f_j^*), \quad \hat{b}_i = \sum_j (\alpha_{ij}^* \hat{a}_j - \beta_{ij}^* \hat{a}_j^\dagger), \quad (43.2)$$

where α and β are the Bogoliubov coefficients. When $\beta_{ij} \neq 0$: the f -vacuum contains g -particles. The expected number of g -particles in the f -vacuum:

$$\langle 0_f | \hat{N}_i^g | 0_f \rangle = \sum_j |\beta_{ij}|^2. \quad (43.3)$$

Particle creation by gravity is encoded in $\beta \neq 0$: the mixing of positive and negative frequencies by the curved background.

43.3 The Unruh effect

An observer accelerating uniformly with proper acceleration a through the Minkowski vacuum $|0_M\rangle$ detects a thermal bath at temperature:

$$T_U = \frac{\hbar a}{2\pi c k_B}. \quad (43.4)$$

The mechanism: the accelerating observer's natural time coordinate (Rindler time) differs from the inertial time. The Bogoliubov transformation between inertial modes (positive frequency in inertial time) and Rindler modes (positive frequency in Rindler time) has $|\beta_{\omega\omega'}|^2/|\alpha_{\omega\omega'}|^2 = e^{-2\pi c\omega/a}$: the Bose–Einstein distribution at temperature T_U .

The Minkowski vacuum, restricted to one Rindler wedge (the accelerating observer's causal diamond), is a thermal density matrix:

$$\hat{\rho}_R = \text{Tr}_L |0_M\rangle \langle 0_M| = \frac{e^{-2\pi c \hat{H}_R/a}}{\text{Tr} e^{-2\pi c \hat{H}_R/a}}, \quad (43.5)$$

where $\hat{H}_R$ is the Rindler Hamiltonian and the trace is over the left Rindler wedge. The thermal character arises from entanglement across the Rindler horizon: the same mechanism as degeneracy (section 9), applied to the bipartition of Minkowski space by the horizon.

43.4 The Hawking effect

A black hole of mass M radiates thermally at the Hawking temperature:

$$T_H = \frac{\hbar c^3}{8\pi G M k_B} = \frac{\hbar \kappa}{2\pi c k_B}, \quad (43.6)$$

where $\kappa = c^4/(4GM)$ is the surface gravity of the Schwarzschild horizon. This is the most profound result in semiclassical gravity: it connects quantum mechanics (the $\hbar$), gravity (the G), thermodynamics (the k_B), and relativity (the c) in a single formula.

The setup. Consider a massless scalar field $\hat{\phi}$ on a spacetime in which a star collapses to form a black hole. Before collapse: the spacetime is approximately flat, and the field is in the Minkowski vacuum $|0_{\text{in}}\rangle$ (no particles at past null infinity $\mathcal{I}^-$). After collapse: the spacetime contains a black hole, and a distant observer at future null infinity $\mathcal{I}^+$ defines “positive frequency” with respect to Schwarzschild time t .

The question: does the distant observer detect particles in the state $|0_{\text{in}}\rangle$? The answer is yes, and the spectrum is thermal.

The mode decomposition. At $\mathcal{I}^-$ (before collapse), expand the field in a complete set of positive-frequency modes $\{f_\omega\}$: $\hat{\phi} = \sum_\omega [f_\omega \hat{a}_\omega + f_\omega^* \hat{a}_\omega^\dagger]$, with $\hat{a}_\omega |0_{\text{in}}\rangle = 0$. At $\mathcal{I}^+$ (after collapse), expand in a different set of modes $\{p_\omega\}$ (positive frequency with respect to Schwarzschild time) and $\{q_\omega\}$ (modes that enter the black hole): $\hat{\phi} = \sum_\omega [p_\omega \hat{b}_\omega + p_\omega^* \hat{b}_\omega^\dagger + q_\omega \hat{c}_\omega + q_\omega^* \hat{c}_\omega^\dagger]$, with $\hat{b}_\omega |0_{\text{out}}\rangle = 0$.

The two sets of modes are related by the Bogoliubov transformation:

$$p_\omega = \int d\omega' [\alpha_{\omega\omega'} f_{\omega'} + \beta_{\omega\omega'} f_{\omega'}^*], \quad (43.7)$$

where $\alpha_{\omega\omega'}$ and $\beta_{\omega\omega'}$ are the Bogoliubov coefficients, computed by propagating the out-modes p_ω backward in time through the collapsing geometry to $\mathcal{I}^-$ and decomposing the result in the in-mode basis.

The key calculation: the Bogoliubov coefficients. The out-modes p_ω at $\mathscr{I}^+$ are positive-frequency with respect to Schwarzschild time t . Near the horizon, the relationship between Schwarzschild time t and the affine parameter U along the horizon generators is exponential:

$$U \sim -e^{-\kappa t/c} \quad (\text{as } t \rightarrow \infty, U \rightarrow 0^-), \quad (43.8)$$

where $\kappa = c^4/(4GM)$ is the surface gravity. This exponential relationship is the origin of everything: it converts the time-translation symmetry $e^{-i\omega t}$ into an exponential mapping $e^{-i\omega \ln(-U)/\kappa}$, which mixes positive and negative frequencies.

A mode of frequency ω at $\mathscr{I}^+$ (positive frequency in t) is traced back through the collapsing geometry. The exponential redshift (43.8) near the horizon maps it onto a superposition of modes at $\mathscr{I}^-$ with both positive and negative frequencies (in U). The negative-frequency component is the β coefficient.

The result (Hawking 1975): the ratio of the Bogoliubov coefficients at frequency ω :

$$\boxed{\frac{|\beta_\omega|^2}{|\alpha_\omega|^2} = e^{-2\pi c\omega/\kappa} = e^{-\hbar\omega/(k_B T_H)}}. \quad (43.9)$$

This is the Boltzmann factor at temperature T_H . The relation $|\alpha|^2 - |\beta|^2 = 1$ (from the canonical commutation relations) then gives:

$$|\beta_\omega|^2 = \frac{1}{e^{\hbar\omega/(k_B T_H)} - 1} = n_{\text{BE}}(\omega, T_H), \quad (43.10)$$

which is the Bose–Einstein distribution at temperature T_H .

The particle number. The expected number of particles detected at $\mathscr{I}^+$ in the in-vacuum $|0_{\text{in}}\rangle$:

$$\langle 0_{\text{in}} | \hat{N}_\omega | 0_{\text{in}} \rangle = |\beta_\omega|^2 = \frac{\Gamma_\omega}{e^{\hbar\omega/(k_B T_H)} - 1}, \quad (43.11)$$

where Γ_ω is the *greybody factor*: the transmission probability of the mode through the effective potential barrier at $r \sim 3r_s/2$. For $\omega \gg \kappa$: $\Gamma_\omega \rightarrow 1$ (high-frequency modes pass freely). For $\omega \lesssim \kappa$: $\Gamma_\omega < 1$ (low-frequency modes are partially reflected by the curvature potential).

The spectrum (43.11) is a perfect thermal spectrum, modified by the greybody factor. The thermal character is exact (it follows from the Bogoliubov coefficient ratio (43.9)); the greybody factor is a grey-body correction that does not affect the temperature, only the overall emission rate.

Numerical scales.

Black hole	M	T_H	Status
Stellar ($10 M_\odot$)	2×10^{31} kg	6×10^{-9} K	Far below CMB (2.7 K); absorbs more than emits
Primordial	$\sim 10^{12}$ kg	$\sim 10^{11}$ K	Evaporating now; endpoint of formation at $t \sim 10^{-23}$ s
Micro (LHC-scale)	$\sim 10^{-23}$ kg	$\sim 10^{26}$ K	Would evaporate in $\sim 10^{-26}$ s; not observed

The physical origin: entanglement. The in-vacuum $|0_{\text{in}}\rangle$ is a pure state of the full quantum field on the collapsing spacetime. When restricted to the exterior of the black hole (tracing over the modes behind the horizon), the reduced density matrix is thermal:

$$\hat{\rho}_{\text{ext}} = \text{Tr}_{\text{int}} |0_{\text{in}}\rangle \langle 0_{\text{in}}| = \frac{e^{-\hat{H}_{\text{ext}}/(k_B T_H)}}{\text{Tr } e^{-\hat{H}_{\text{ext}}/(k_B T_H)}}. \quad (43.12)$$

The Hawking radiation is *not* produced by the black hole “from the inside.” It arises because the in-vacuum, viewed from the exterior alone, is an entangled state: the modes inside and outside the horizon are correlated, and tracing over the interior produces a thermal density matrix. This is the same mechanism as degeneracy (section 9): restricting a pure state to a subsystem produces a mixed state, and the degree of mixedness is determined by the entanglement between the subsystem and its complement.

In the Fock-Born language: the Fock hidden state is $|0_{\text{in}}\rangle$ (pure, one vector, rotating). The Born layer renders this state from the perspective of the exterior observer, who has access only to the exterior degrees of freedom. The Born distribution is thermal because the exterior Born-layer observable ($\hat{N}_\omega$ restricted to $\mathcal{S}^+$) has enormous degeneracy: many interior configurations produce the same exterior particle number. The temperature T_H is the degeneracy parameter.

43.5 Black hole evaporation

The Hawking radiation carries energy away from the black hole. By energy conservation, the black hole loses mass: $dMc^2/dt = -P_H$, where P_H is the radiated power.

The luminosity. The Stefan–Boltzmann law for the Hawking radiation, summing over all species and including greybody factors:

$$P_H = \sigma_{\text{eff}} A T_H^4 \sim \frac{\hbar c^6}{G^2 M^2}, \quad (43.13)$$

where $A = 16\pi G^2 M^2/c^4$ is the horizon area, $T_H = \hbar c^3/(8\pi G M k_B)$, and σ_{eff} includes the greybody factor and the number of light species. The luminosity scales as M^{-2} : lighter black holes radiate faster.

The lifetime. Since $P_H \propto M^{-2}$ and $E = Mc^2$: $dM/dt \propto -M^{-2}$, giving $M^3(t) = M_0^3 - \alpha t$ with $\alpha = 3\sigma_{\text{eff}}\hbar c^4/(256\pi^3 G^2)$. The evaporation time:

$$t_{\text{evap}} = \frac{5120\pi G^2 M_0^3}{\hbar c^4} \approx 2 \times 10^{67} \text{ yr} \left(\frac{M_0}{M_\odot} \right)^3. \quad (43.14)$$

A solar-mass black hole evaporates in $\sim 10^{67}$ years — far longer than the age of the universe ($\sim 10^{10}$ years). A primordial black hole with $M_0 \sim 5 \times 10^{11}$ kg would be evaporating today ($t_{\text{evap}} \sim 10^{10}$ years). A Planck-mass black hole ($M_P \sim 10^{-8}$ kg) evaporates in $\sim t_P \sim 10^{-44}$ s.

The endpoint. As $M \rightarrow 0$: $T_H \rightarrow \infty$ and $P_H \rightarrow \infty$. The final moments of evaporation are explosive: the black hole radiates all SM species (including heavy quarks, W , Z , Higgs) in a burst of $\sim M_0 c^2$ energy. The semiclassical calculation breaks down when $M \lesssim M_P$ (the Planck mass), where quantum gravity effects become important. The endpoint — whether a Planck-mass remnant, a final explosion, or something else — is unknown.

The negative specific heat. The heat capacity of a Schwarzschild black hole:

$$C = \frac{dE}{dT_H} = \frac{d(Mc^2)}{dT_H} = -\frac{8\pi G M^2 k_B}{c\hbar} = -\frac{k_B c^3 A}{16\pi G \hbar} < 0. \quad (43.15)$$

The specific heat is negative: as the black hole loses energy (by radiating), its temperature *increases*. Hotter $\rightarrow$ radiates faster $\rightarrow$ loses mass faster $\rightarrow$ gets hotter: a runaway process. Black holes are thermodynamically unstable in the canonical ensemble.

This is the opposite of normal thermodynamic systems (where losing energy cools the system, stabilising the equilibrium). The negative specific heat is a consequence of the inverse relationship $T_H \propto 1/M$: smaller black holes are hotter.

43.6 The information paradox

The Hawking radiation is thermal: the density matrix $\hat{\rho}_{\text{ext}} = e^{-\hat{H}/(k_B T_H)}/Z$ is a maximum-entropy state for the given energy. The thermal state contains no information about the initial state of the matter that formed the black hole (except M, J, Q). If the black hole evaporates completely, the final state is a thermal bath of Hawking quanta — a mixed state — even though the initial state (the collapsing star) was pure.

This is the *information paradox* (Hawking, 1976): the evolution from a pure initial state to a mixed final state violates unitarity. The S -matrix is not unitary: $S^\dagger S \neq 1$.

The paradox stated precisely. Three premises, each well-motivated, are mutually inconsistent:

Premise 1 (unitarity): quantum evolution is unitary. The full state of the universe evolves as $|\Psi(t)\rangle = e^{-iHt} |\Psi(0)\rangle$, and the evolution operator e^{-iHt} is unitary. A pure state remains pure.

Premise 2 (Hawking’s calculation): the radiation is thermal. The Bogoliubov calculation, performed on a fixed Schwarzschild background, gives a thermal spectrum (43.11) that carries no information about the initial state beyond M .

Premise 3 (no remnant): the black hole evaporates completely. No Planck-mass remnant is left behind that could store the missing information.

If all three hold: the initial pure state evolves to a final mixed state — unitarity is violated. At least one premise must be modified.

The Page curve. Page (1993) argued that if unitarity is preserved, the entanglement entropy of the Hawking radiation must follow a specific curve: initially increasing (as entangled Hawking pairs are emitted, one to infinity and one behind the horizon) and then decreasing (as the black hole shrinks and the radiation purifies). The turnover occurs at the *Page time* $t_P \sim t_{\text{evap}}/2$ (when the black hole has lost half its initial entropy).

The Page curve is a Born-layer prediction: it describes the entanglement entropy of the radiation as a function of time, which in principle could be measured by collecting all Hawking quanta and performing quantum state tomography. Hawking’s calculation gives a monotonically increasing entropy (no turnover): the radiation never purifies. The discrepancy between the unitary prediction (Page curve) and the semiclassical prediction (monotonically increasing) is the information paradox expressed as a curve.

Recent developments. The island formula (Penington, Almheiri, Marolf, Polchinski, Sully, 2019–2020) modifies the entanglement entropy calculation by including “island” contributions — regions behind the horizon that are included in the entropy calculation of the radiation via the quantum extremal surface prescription. The island formula reproduces the Page curve: the entanglement entropy of the radiation increases, turns over at the Page time, and decreases to zero as the black hole evaporates. Unitarity is preserved.

The island formula has been derived rigorously in specific models (Jackiw–Teitelboim gravity coupled to a CFT, and in the Penington–Shenker–Stanford–Yang framework using replica wormholes). Whether it resolves the paradox in realistic four-dimensional gravity is an open question.

The Fock-Born reading. The information paradox is a tension between two Fock-Born principles:

Fock-layer unitarity: the Schrödinger equation $\dot{h}_t = -iHh_t$ is unitary. A pure h_0 evolves to a pure h_t . No information is lost in the Fock layer.

Born-layer thermality: the exterior observer, performing Born measurements on the radiation, obtains thermal statistics. The Born distribution $P(\lambda) = \text{Tr}_{\text{int}}[P_\lambda |0_{\text{in}}\rangle \langle 0_{\text{in}}| P_\lambda]$ is thermal because the observer has access only to exterior degrees of freedom.

These are not contradictory: a pure Fock-layer state can produce a thermal Born distribution when the observable has large degeneracy (section 9). The apparent paradox arises from asking whether the Fock-layer state is recovered after the black hole evaporates. If the interior degrees of freedom are inaccessible (behind the horizon, then destroyed at the singularity), the entanglement is lost and the Fock-layer state cannot be reconstructed from the exterior alone.

The resolution, in the Fock-Born framework, depends on whether the interior Fock-layer degrees of freedom survive the evaporation. If they do (via the island mechanism, or a remnant, or some other process): unitarity is preserved and the information is encoded in subtle correlations among the Hawking quanta. If they do not: unitarity is violated and the Fock-Born architecture requires modification at the Planck scale.

The information paradox is the sharpest test of the Fock-Born architecture in the gravitational regime. Its resolution will determine whether the one-way valve (section 43.15) — the prohibition on recycling Born-layer outputs into the Fock layer — extends to the gravitational sector, or whether gravity requires a fundamentally different architecture.

43.7 The Bekenstein–Hawking entropy and the heat kernel

The Hawking temperature gives a thermodynamic entropy via $dS = dE/T$, with $E = Mc^2$ and $T = T_H$:

$$S_{\text{BH}} = \frac{k_B c^3}{4\hbar G} A = \frac{k_B A}{4\ell_P^2}, \quad (43.16)$$

where $A = 16\pi G^2 M^2/c^4$ is the horizon area and $\ell_P = \sqrt{\hbar G/c^3}$ is the Planck length. The entropy is proportional to the area in Planck units: one bit per four Planck areas.

The heat kernel derivation. The one-loop effective action for a quantum field on the Schwarzschild background is $\Gamma^{(1)} = -\frac{1}{2} \int_0^\infty \frac{ds}{s} \text{Tr} e^{-sD}$ (section 23.11). The trace of the heat kernel on a manifold with boundary (the horizon) receives a boundary contribution from the conical singularity at the Euclidean horizon. The Euclidean Schwarzschild metric, periodically identified in imaginary time with period $\beta = 1/(k_B T_H)$, has a conical singularity at $r = r_s$ unless $\beta = 8\pi G M/c^3$ —which is exactly the inverse Hawking temperature. The a_1 coefficient of the heat kernel on the conical geometry gives a contribution $\propto A/\varepsilon^2$ (where ε is the UV cutoff), which reproduces the area-entropy relation (43.16) when $\varepsilon \sim \ell_P$.

The heat kernel thus connects the Hawking temperature (from the Bogoliubov transformation), the Bekenstein–Hawking entropy (from thermodynamics), and the Seeley–DeWitt coefficients (from the UV structure of the operator D). All three are computed from the same object: e^{-sD} on the Euclidean Schwarzschild background.

43.8 The semiclassical synthesis

The semiclassical framework—classical $g_{\mu\nu}$, quantum $\hat{\psi}$ —delivers:

Result	Input	Mechanism
Unruh effect	$\eta_{\mu\nu}$, acceleration a	Bogoliubov: Rindler/Minkowski mode mixing
Hawking radiation	Schwarzschild $g_{\mu\nu}$	Bogoliubov: in/out mode mixing across horizon
BH entropy	T_H , thermodynamics	$S_{\text{BH}} = k_B A/(4\ell_P^2)$
Trace anomaly	Any $g_{\mu\nu}$	$\langle T^\mu{}_\mu \rangle = a_2/(16\pi^2)$: Seeley–DeWitt

All of these follow from the action principle and the heat kernel, without quantising $g_{\mu\nu}$ itself. The semiclassical level is remarkably powerful: it connects quantum field theory, thermodynamics, and geometry through the Bogoliubov transformation and the Seeley–DeWitt expansion.

Part II: Quantum $g_{\mu\nu}$ — existing approaches

43.9 The graviton: a spin-2 field on $\eta_{\mu\nu}$

In the weak-field limit, write $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}/M_P$, where $h_{\mu\nu}$ is a small perturbation and $M_P = \sqrt{\hbar c/(8\pi G)}$ is the reduced Planck mass. Expanding the Einstein–Hilbert action to quadratic order gives the Fierz–Pauli action for a free massless spin-2 field:

$$S_{\text{FP}} = \int d^4x \left[-\frac{1}{4}(\partial_\rho h_{\mu\nu})^2 + \frac{1}{4}(\partial_\mu h)^2 + \frac{1}{2}(\partial^\mu h_{\mu\nu})^2 - \frac{1}{2}(\partial_\mu h)(\partial^\nu h^\mu{}_\nu) \right]. \quad (43.17)$$

The graviton has spin 2, helicities $\lambda = \pm 2$, and couples universally to all energy-momentum via $S_{\text{int}} \supset (1/M_P) \int h_{\mu\nu} T^{\mu\nu}$.

Weinberg’s soft graviton theorem. Weinberg (1964) showed that consistency of the S -matrix for massless spin-2 particles requires that the coupling be universal and proportional to $T_{\mu\nu}$ —and that the low-energy scattering amplitudes reproduce the Einstein equation. General relativity is the unique low-energy theory of a massless spin-2 particle. This is GR derived from QFT, not the other way around.

43.10 Effective field theory of gravity

Weinberg’s effective field theory programme (section ??) applies to gravity: write the most general Lagrangian consistent with diffeomorphism invariance, organised by the number of derivatives:

$$\mathcal{L}_{\text{grav}} = \frac{M_P^2}{2} R + c_1 R^2 + c_2 R_{\mu\nu} R^{\mu\nu} + c_3 R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma} + \dots \quad (43.18)$$

The Einstein–Hilbert term (R , two derivatives) dominates at low energies. Higher-curvature terms (R^2 , four derivatives) are suppressed by E^2/M_P^2 and contribute $o(a/\lambda)$ corrections.

At one loop: the graviton self-energy is finite (the R^2 counterterm has the right form). At two loops: a counterterm $R_{\mu\nu\rho\sigma} R^{\rho\sigma\alpha\beta} R_{\alpha\beta}{}^{\mu\nu}$ is required, with a coefficient that cannot be absorbed into the existing parameters. Gravity is non-renormalisable: each loop order requires new counterterms with new coefficients.

In the lattice-first viewpoint: non-renormalisability means the lattice Hamiltonian $H_{\text{grav}}(a)$ has infinitely many parameters. This is not a sickness; it means gravity is a many-parameter effective theory, predictive at $E \ll M_P$ but requiring the full lattice specification at $E \sim M_P$.

43.11 The problem of time

Canonical quantisation of gravity (Dirac, Arnowitt–Deser–Misner, DeWitt, Wheeler) leads to the Wheeler–DeWitt equation:

$$\hat{H} |\Psi\rangle = 0, \quad (43.19)$$

where $\hat{H}$ is the Hamiltonian constraint. The wavefunction $\Psi[h_{ij}]$ is a functional of the spatial metric h_{ij} on a three-surface. The equation has no time variable: $\hat{H} = 0$ says the total energy is zero (as required by diffeomorphism invariance). Time does not appear in the fundamental equation.

This is the *problem of time*: the Schrödinger equation $\dot{h}_t = -iHh_t$ requires a Hamiltonian that generates time evolution, but the Hamiltonian constraint $H = 0$ generates nothing. The Fock-Born architecture requires a non-trivial Hamiltonian (the engine must run); the Wheeler–DeWitt equation gives a frozen engine.

The problem arises because quantising $g_{\mu\nu}$ quantises the time coordinate itself. In the Fock-Born framework: time is part of the Fock-layer substrate (the lattice has a time direction with spacing a_t), not a Born-layer variable to be quantised. Quantising time is recycling the engine’s clock into the engine’s state—a category error.

43.12 Existing approaches, examined from Fock-Born

Canonical quantum gravity / loop quantum gravity. LQG quantises the spatial geometry directly: the Hilbert space is spanned by spin networks, and area and volume operators have discrete spectra. In the Fock-Born reading: LQG attempts to build a Fock-layer engine whose state space encodes the spatial geometry. The discrete spectra of area and volume are consistent with a lattice substrate. But the problem of time remains: the Hamiltonian constraint $H = 0$ leaves the engine frozen. LQG has not produced a dynamics (a non-trivial H) that reproduces the semiclassical limit.

String theory. String theory replaces point particles with one-dimensional strings propagating on a fixed background $g_{\mu\nu}$. The graviton emerges as a massless spin-2 excitation of the closed string. The theory is UV finite (no divergences at any loop order) and contains gravity automatically. In the Fock-Born reading: string theory provides a specific Fock-layer engine (the string worldsheet) on a fixed background. But the background $g_{\mu\nu}$ is an input, not an output—the theory does not derive $g_{\mu\nu}$ from a more fundamental substrate. The landscape of $\sim 10^{500}$ string vacua suggests that the theory has many parameters (non-renormalisable in the Weinberg sense), consistent with the lattice-first expectation.

Asymptotic safety. The asymptotic safety programme (Weinberg 1979, Reuter 1998) seeks a non-trivial UV fixed point of the gravitational RG flow, which would make gravity non-perturbatively renormalisable with finitely many parameters. In the Fock-Born reading: asymptotic safety asks whether the rendering-stability condition can be satisfied with finitely many $\theta(a)$ for all $a \rightarrow 0$. The lattice-first viewpoint regards this as a question about the rendering, not about the engine: the engine is defined at finite a , and the continuum limit $a \rightarrow 0$ is a mathematical convenience. Whether the fixed point exists is interesting but not physically necessary.

Induced gravity. Sakharov (1967) observed that integrating out matter fields on a curved background generates an effective action for $g_{\mu\nu}$ that includes the Einstein–Hilbert term: $\Gamma_{\text{ind}}[g] \supset \frac{1}{16\pi G_{\text{ind}}} \int \sqrt{-g} R$, with G_{ind} determined by the matter content. The gravitational action is not postulated; it is *induced* by quantum matter. In the Fock-Born reading: the regularizer ($\int \sqrt{-g} R$) is not a separate postulate but a one-loop consequence of the matter action—the heat kernel’s a_1 coefficient on a curved background. This supports the view that $g_{\mu\nu}$ is a Born-layer output derived from matter fields on a flat substrate.

Jacobson’s thermodynamic derivation. Jacobson (1995) derived the Einstein equation from the thermodynamic relation $\delta Q = T dS$ applied to local causal horizons, using the Unruh

temperature and the Bekenstein–Hawking area-entropy relation. In the Fock-Born reading: Jacobson’s derivation is a Born-layer argument—it derives the equation of state for $g_{\mu\nu}$ from the thermodynamic properties of the rendering, without specifying the Fock-layer engine. This is consistent with the view that the Einstein equation is a macroscopic equation for a Born-layer output, not a fundamental equation for a Fock-layer variable.

43.13 Induced gravity: the Einstein–Hilbert action from the heat kernel

Sakharov (1967) made a remarkable observation: the Einstein–Hilbert action $\frac{1}{16\pi G} \int \sqrt{-g} R$ need not be postulated. It is *induced* by quantum matter fields propagating on a curved background. Integrating out the matter produces an effective action for $g_{\mu\nu}$ whose leading terms are exactly the cosmological constant and the Einstein–Hilbert action, with the gravitational constant G determined by the matter content.

This is the deepest application of the heat kernel in this book: it derives gravity from quantum field theory on a flat substrate, without ever postulating a gravitational action.

The curved-background operator

Consider a free scalar field ϕ of mass m on a curved background $g_{\mu\nu}$ with non-minimal coupling ξ to the scalar curvature:

$$S[\phi; g] = -\frac{1}{2} \int d^4x \sqrt{g} \phi D \phi, \quad D = -\nabla^2 + m^2 + \xi R, \quad (43.20)$$

where $\nabla^2 = g^{\mu\nu} \nabla_\mu \nabla_\nu$ is the covariant Laplacian and R is the Ricci scalar. The operator D is a second-order elliptic differential operator on the Riemannian manifold (M, g) .

The one-loop effective action for $g_{\mu\nu}$, obtained by integrating out ϕ :

$$\Gamma^{(1)}[g] = \frac{1}{2} \ln \det D = \frac{1}{2} \text{Tr} \ln D. \quad (43.21)$$

The trace is over the Hilbert space $L^2(M, \sqrt{g} d^4x)$ of square-integrable functions on the curved manifold. This is a functional of $g_{\mu\nu}$: the metric enters through the covariant Laplacian ∇^2 , the scalar curvature R , and the volume element $\sqrt{g}$.

The heat kernel on a curved manifold

The Schwinger proper-time representation:

$$\Gamma^{(1)}[g] = -\frac{1}{2} \int_\varepsilon^\infty \frac{ds}{s} \text{Tr} e^{-sD} = -\frac{1}{2} \int_\varepsilon^\infty \frac{ds}{s} K(s), \quad (43.22)$$

where $\varepsilon \sim a^2$ is the proper-time cutoff (related to the lattice spacing), and $K(s) = \text{Tr} e^{-sD} = \int d^4x \sqrt{g} K(x, x; s)$ is the integrated heat kernel. The diagonal heat kernel $K(x, x; s)$ satisfies the heat equation $(\partial_s + D_x)K(x, y; s) = 0$ with initial condition $K(x, y; 0) = \delta^4(x - y)/\sqrt{g}$.

The key technical result is the *Seeley–DeWitt asymptotic expansion*: for small s (short proper time, high energy),

$$K(x, x; s) = \frac{1}{(4\pi s)^{d/2}} \sum_{n=0}^{\infty} a_n(x) s^n, \quad (43.23)$$

where d is the spacetime dimension and $a_n(x)$ are the *Seeley–DeWitt coefficients*—local scalar invariants built from the curvature $R_{\mu\nu\rho\sigma}$, the endomorphism $E = m^2 + \xi R$, and their covariant derivatives.

The expansion is universal: it depends only on the local geometry near x and the structure of D , not on the global topology of M or the boundary conditions. This is why the heat kernel is the right tool: it computes the UV structure of the effective action purely from local data.

The Seeley–DeWitt coefficients for the curved Laplacian

For the operator $D = -\nabla^2 + E$ with $E = m^2 + \xi R$ on a Riemannian manifold of dimension $d = 4$, the first three coefficients are:

$a_0(x) = 1$. The zeroth coefficient is trivial: it counts the number of field components (one for a real scalar). For a field with n_s internal degrees of freedom (e.g., $n_s = 4$ for a Dirac fermion in $d = 4$): $a_0 = n_s$.

$a_1(x) = (\frac{1}{6} - \xi)R - m^2$. The first coefficient involves the scalar curvature R and the mass. For conformal coupling $\xi = 1/6$: the curvature term vanishes and $a_1 = -m^2$. For minimal coupling $\xi = 0$: $a_1 = R/6 - m^2$.

The derivation of a_1 proceeds from the short-distance expansion of the heat kernel. Write $K(x, y; s) = \frac{1}{(4\pi s)^{d/2}} e^{-\sigma(x, y)/(2s)} \Omega(x, y; s)$, where $\sigma(x, y) = \frac{1}{2}d^2(x, y)$ is half the squared geodesic distance and Ω is a smooth bitensor. Substituting into the heat equation $(\partial_s + D_x)K = 0$ and expanding $\Omega(x, y; s) = \sum_n \Omega_n(x, y) s^n$: the coefficients Ω_n satisfy a recursion relation. At coincidence $y \rightarrow x$: $\Omega_0(x, x) = 1$ and $\Omega_1(x, x) = (\frac{1}{6} - \xi)R - m^2$. The recursion uses the expansion of the geodesic distance $\sigma(x, y)$ and the Van Vleck–Morette determinant $\Delta(x, y) = -\det(-\nabla_\mu \nabla'_\nu \sigma) / \sqrt{g(x)g(y)}$ in Riemann normal coordinates. The result $a_1 = \Omega_1(x, x)$ follows after collecting the terms involving $R_{\mu\nu\rho\sigma}$ contracted to scalars; at first order, only R contributes.

$a_2(x)$ for a general operator. For $D = -(g^{\mu\nu} \nabla_\mu \nabla_\nu + E)$ acting on sections of a vector bundle with connection ω_μ (curvature $\Omega_{\mu\nu} = \partial_\mu \omega_\nu - \partial_\nu \omega_\mu + [\omega_\mu, \omega_\nu]$):

$$a_2(x) = \frac{1}{180}(R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} - R_{\mu\nu}R^{\mu\nu} + \square R) + \frac{1}{2}(E - \frac{1}{6}R)^2 + \frac{1}{12}\Omega_{\mu\nu}\Omega^{\mu\nu} + \frac{1}{6}\square E. \quad (43.24)$$

For a real scalar with $E = m^2 + \xi R$ and $\Omega_{\mu\nu} = 0$ (no internal gauge field): the Ω^2 term vanishes, and the remaining terms are polynomial in R , $R_{\mu\nu}$, $R_{\mu\nu\rho\sigma}$, and m .

The a_2 coefficient is the payoff: it determines the one-loop beta functions of all couplings (as exploited in section 39.8 for the Standard Model gauge couplings), and on a curved background it determines the gravitational counterterms.

The induced effective action

Substituting the Seeley–DeWitt expansion (43.23) into the proper-time integral (43.22) and integrating over s term by term:

$$\begin{aligned} \Gamma^{(1)}[g] &= -\frac{1}{2} \int_{\varepsilon}^{\infty} \frac{ds}{s} \frac{1}{(4\pi s)^2} \int d^4x \sqrt{g} [a_0 + a_1 s + a_2 s^2 + \dots] \\ &= -\frac{1}{2(4\pi)^2} \int d^4x \sqrt{g} \left[\frac{a_0}{2\varepsilon^2} + \frac{a_1}{\varepsilon} + a_2 \ln \frac{1}{\varepsilon\mu^2} + \text{finite} \right]. \end{aligned} \quad (43.25)$$

The three divergent terms (in $d = 4$) are:

The a_0 term: $\frac{a_0}{2\varepsilon^2} \int \sqrt{g}$. This is a cosmological constant: a term proportional to the spacetime volume $\int \sqrt{g} d^4x$. With $\varepsilon \sim a^2$ (the proper-time cutoff related to the lattice spacing):

$$\Gamma^{(1)} \supset -\frac{1}{2(4\pi)^2} \cdot \frac{1}{2a^4} \int d^4x \sqrt{g} = -\frac{\Lambda_{\text{ind}}}{16\pi G_{\text{ind}}} \int d^4x \sqrt{g}, \quad (43.26)$$

which is a contribution to the cosmological constant from each matter species. This is the $\mathbf{A}_0$ term of section 39: the vacuum energy, quartic in the cutoff.

The a_1 term: $\frac{a_1}{\varepsilon} \int \sqrt{g}$. With $a_1 = (\frac{1}{6} - \xi)R - m^2$ and $\varepsilon \sim a^2$:

$$\Gamma^{(1)} \supset -\frac{1}{2(4\pi)^2} \cdot \frac{1}{a^2} \int d^4x \sqrt{g} \left[\left(\frac{1}{6} - \xi \right) R - m^2 \right]. \quad (43.27)$$

The R -dependent piece is the *induced Einstein–Hilbert action*:

$$\Gamma_{\text{ind}}^{(1)} \supset \frac{1}{16\pi G_{\text{ind}}} \int d^4x \sqrt{g} R, \quad \frac{1}{G_{\text{ind}}} = \frac{(\frac{1}{6} - \xi)}{2\pi a^2} \times (\text{per species}). \quad (43.28)$$

The gravitational constant G is not a free parameter of the fundamental theory: it is *generated* by the one-loop correction from matter fields on the curved background. Each matter species contributes to $1/G_{\text{ind}}$ with a coefficient that depends on its spin, mass, and non-minimal coupling ξ .

For minimally coupled fields ($\xi = 0$): all matter species contribute positively to $1/G_{\text{ind}}$, giving an attractive gravitational force. The total induced Newton constant, summing over all Standard Model fields:

$$\frac{1}{G_{\text{ind}}} = \frac{1}{12\pi a^2} \sum_i \eta_i n_i \left(\frac{1}{6} - \xi_i \right), \quad (43.29)$$

where $\eta_i = +1$ for bosons, -1 for fermions, n_i is the number of components, and ξ_i is the non-minimal coupling (typically $\xi_i = 0$ for all SM fields, giving $1/G_{\text{ind}} \propto 1/a^2$, i.e., $G_{\text{ind}} \propto a^2$).

The m^2 piece in a_1 gives a quadratic contribution to the cosmological constant (the $\mathbf{A}_1$ term of section 39).

The a_2 term: $a_2 \ln(1/\varepsilon\mu^2) \int \sqrt{g}$. This gives logarithmically divergent corrections to higher-curvature terms (R^2 , $R_{\mu\nu}R^{\mu\nu}$, etc.) in the effective action. These are the gravitational counterterms that make gravity non-renormalisable: at each loop order, new curvature invariants appear that cannot be absorbed into the existing parameters. In the effective field theory framework (section ??), they are the $O(\alpha'^2)$ corrections suppressed by $(E/M_P)^2$.

Summing over species

For the full Standard Model, the induced gravitational constant receives contributions from all species:

$$\frac{1}{G_{\text{ind}}} = \frac{1}{12\pi a^2} \left[\underbrace{28 \times \frac{1}{6}}_{\text{bosons}} - \underbrace{90 \times \frac{1}{12}}_{\text{fermions}} \right] = \frac{1}{12\pi a^2} \left[\frac{28}{6} - \frac{90}{12} \right] = -\frac{1}{12\pi a^2} \cdot \frac{17}{6}, \quad (43.30)$$

where the 28 bosonic and 90 fermionic helicity states of the SM are counted with their respective contributions (1/6 per real scalar, 1/12 per Weyl fermion component for minimal coupling). The negative sign means the fermion contribution dominates: the induced G is positive (attractive gravity), as observed.

Setting $G_{\text{ind}} = G_{\text{Newton}}$:

$$a \sim \sqrt{\frac{17}{72\pi G_{\text{Newton}}}} \sim \ell_P \times O(1), \quad (43.31)$$

giving a lattice spacing of order the Planck length. The gravitational constant, in the induced-gravity picture, is not an independent input: it is determined by the matter content and the lattice spacing. Newton's constant is a derived quantity.

The Fock-Born reading

Induced gravity is the strongest version of the Fock-Born thesis applied to gravity:

The Fock layer is a lattice with spacing a , containing the Standard Model fields as matter degrees of freedom. The lattice is flat: there is no dynamical metric in the Fock layer. The Hamiltonian $H_{\text{SM}}(a)$ is constructed on the flat lattice, with no gravitational terms.

The Born layer renders the metric $g_{\mu\nu}$ as a macroscopic classical field at $\lambda \gg a$. The one-loop effective action for $g_{\mu\nu}$ —obtained by integrating out the matter fields on the curved background via the heat kernel—generates the Einstein–Hilbert action with Newton’s constant determined by the matter content. The Einstein equation is not postulated; it is a one-loop consequence of matter quantum mechanics on a curved background.

The gravitational action is the a_1 coefficient. The regulariser $\int \sqrt{g} R$ that appears in the action principle for general relativity is not an independent postulate. It is the a_1 Seeley–DeWitt coefficient of the matter heat kernel, integrated over proper time. The “curvature cost” in the corrected Wheeler dictum (section 42.6) is generated by quantum matter, not by a separate gravitational sector.

The cosmological constant is the a_0 coefficient. The vacuum energy term $\int \sqrt{g} \Lambda$ is the a_0 Seeley–DeWitt coefficient: the count of degrees of freedom, weighted by spin statistics. Its fine-tuning to $\sim 10^{-123}$ of its natural value is the most extreme instance of the rendering-stability condition.

Newton’s constant sets the lattice spacing. If all of G_{Newton} is induced by SM matter loops, then $G \propto a^2$ and the lattice spacing is $a \sim \ell_P$. The Planck scale is not a new energy scale introduced by gravity; it is the lattice spacing of the Fock-layer engine, determined by the requirement that the induced Newton constant matches the observed value.

This does not resolve the problem of quantum gravity (the problem of time, the non-renormalisability at two loops, the information paradox remain), but it shifts the framing: gravity is not a fundamental force to be quantised. It is an emergent interaction, induced by quantum matter on a flat substrate, whose classical equation of state (the Einstein equation) is the a_1 coefficient of the heat kernel rendered at macroscopic scales.

43.14 Bare-parameter sensitivity and asymptotic safety

The rendering-stability condition $\mathcal{O}_{\text{Born}} = \mathcal{O}_{\text{phys}} + o(a/\lambda)$ requires the Born-layer outputs to be a -independent. It says nothing about the bare parameters $\theta(a)$. These are free to vary as steeply as they like, provided the variations conspire to produce a stable rendering.

We now classify the bare parameters by their sensitivity to a and identify which concerns are genuine.

The sensitivity exponent θ

Each bare parameter approaches its rendered value at a rate controlled by a *sensitivity exponent* θ :

$$\theta(a) - \theta_{\text{phys}} \propto a^{-\theta}, \quad (43.32)$$

where $\theta > 0$ for a relevant coupling (the bare parameter diverges as $a \rightarrow 0$) and $\theta = 0$ for a marginal coupling (logarithmic variation). A factor-of-2 change in a changes the bare parameter by a factor of 2^θ : the larger θ , the steeper the calibration curve.

θ	Variation	Factor-of-2 in a	Example	Stability
0	$\ln(1/a)$	$+\ln 2$ shift	α_s, m_e	Gentle (logarithmic)
1	$1/a$	$\times 2$	(hypothetical)	Moderate
2	$1/a^2$	$\times 4$	m_H^2	Steep (30-digit cancel.)
4	$1/a^4$	$\times 16$	Λ (traditional)	Extreme (120 digits)

The numerical stability of the lattice engine degrades with θ : larger θ means the bare parameter is more sensitive to the lattice spacing, requiring more precise cancellation to produce the rendered output.

Which couplings are bare parameters?

Not every coupling in the effective action is a bare parameter of the engine. In the induced gravity framework (section 43.13), the Fock-layer engine is $H_{\text{SM}}(a)$ on a flat lattice with no independent gravitational sector. The gravitational couplings are Born-layer outputs, computed from the matter content via the heat kernel:

Newton’s constant (G): the a_1 Seeley–DeWitt coefficient. $G(a) \propto a^2/N_{\text{species}}$. This is not a parameter that someone dials on the engine. It is the Born-layer rendering of the matter vacuum polarisation on a curved background. Its “running” is a property of the heat kernel, not a calibration of the engine.

The cosmological constant (Λ): the a_0 Seeley–DeWitt coefficient. $\rho_\Lambda \propto N_{\text{species}}/a^4$. Again: not a bare parameter, but a Born-layer output.

In the traditional treatment (where gravity has its own bare action): $\Lambda_{\text{phys}} = \Lambda_0 + \delta\Lambda$, with Λ_0 a bare parameter and $\delta\Lambda \sim 1/a^4$ the matter-loop correction. The cancellation of 120 digits between Λ_0 and $\delta\Lambda$ is the cosmological constant “problem.”

In the induced framework: there is no Λ_0 . The engine has no gravitational bare parameter. The cosmological constant is a single computed number — the vacuum energy of the matter fields, rendered to the Born layer. There is nothing to cancel against. The “problem” is an artefact of treating a Born-layer output as a bare parameter.

The corrected classification:

Coupling	Layer	Variation with a	θ	Concern?
α_s	Fock (bare)	$\sim \ln(1/a)$	0	None (gentle)
m_e	Fock (bare)	$\sim m_e \ln(1/a)$	0	None (chiral)
m_H^2	Fock (bare)	$\sim 1/a^2$	2	Real (no protection)
G	Born (induced, a_1)	$\sim a^2/N$	—	None (not a bare param.)
Λ	Born (induced, a_0)	$\sim 1/a^4$	—	None (not a bare param.)

The only Fock-layer bare parameter with a genuine sensitivity concern is m_H^2 . It runs as $1/a^2$, with no symmetry protection (section 29.6: the Wald identity protects gauge bosons; chiral symmetry protects fermions; nothing protects scalars). Newton’s constant and the cosmological constant, in the induced framework, are Born-layer outputs and carry no bare-parameter

sensitivity.

The Higgs mass: the one real concern

The Higgs bare mass at lattice spacing a :

$$m_H^2(a) = m_{H,\text{phys}}^2 - \frac{c\pi^2}{a^2}, \quad c = \frac{1}{16\pi^2} [-6y_t^2 + \frac{9}{4}g^2 + \frac{3}{4}g'^2 + 6\lambda]. \quad (43.33)$$

If $a \sim \ell_P$: $m_H^2(a) \sim -10^{34} \text{ GeV}^2$, and the rendered output $(125 \text{ GeV})^2$ requires a cancellation of 30 decimal places between $m_H^2(a)$ and the loop correction.

This is a computational-substrate concern: the engine's calibration is numerically sensitive. A change in a of one part in 10^{15} changes $m_H^2(a)$ by $O(1) \text{ GeV}^2$, which changes the rendered Higgs mass by a detectable amount. The engine must be calibrated to extraordinary precision.

Possible resolutions.

A symmetry protects m_H^2 : supersymmetry pairs each boson with a fermion, cancelling the $1/a^2$ coefficient: $c_{\text{SUSY}} = 0$, and the correction becomes $\delta m_H^2 \sim m_{\text{soft}}^2 \ln(1/a)$ (logarithmic, $\theta = 0$). The bare parameter becomes gentle. This requires superpartners at $m_{\text{soft}} \sim \text{TeV}$ — not yet observed.

The Higgs is a pNGB: if m_H is protected by an approximate symmetry (composite Higgs models, the bare mass is $m_H^2 \propto g^2 f^2 / (16\pi^2)$ with $f \sim \text{TeV}$: naturally small, $\theta \approx 0$. Requires compositeness at the TeV scale — not yet observed.

The lattice spacing is not Planckian: if $a \sim 1/\text{TeV}$ rather than $a \sim \ell_P$, the cancellation is $O(1)$: no fine-tuning. But then what sets a ?

It is simply a fact: the rendering-stability condition determines $m_H^2(a)$ as a function of a , and the function is steep. The engine is calibrated this way. No explanation needed, no inconsistency.

What asymptotic safety adds

Weinberg's asymptotic safety programme (1979) asks whether the dimensionless couplings $\tilde{g}_i(a) = g_i(a) \cdot a^{-[g_i]}$ approach a non-trivial fixed point $\tilde{g}_i^*$ as $a \rightarrow 0$.

If the fixed point exists with n_{rel} relevant directions: the trajectory $\theta(a)$ is determined by n_{rel} parameters. Every bare coupling — including $m_H^2(a)$ — is a function of those parameters. The steep calibration of $m_H^2(a)$ is not a free choice; it is a consequence of the fixed-point trajectory. The number 30 (the number of cancelled digits) is determined by $\tilde{g}^*$ and θ , not by an independent fine-tuning.

If the fixed point does not exist: each bare coupling is independent. The steep calibration is a brute fact.

In either case: the rendered physics is the same at $E \ll M_P$. The fixed point determines the *economy of description* (how many inputs the engine needs), not the *validity of predictions* (what the engine outputs).

But the fixed point does not reduce θ . Even if a fixed point exists: the sensitivity exponent θ of each relevant coupling is determined by the fixed-point structure. The fRG calculations find $\theta \sim 2\text{--}4$ for the gravitational couplings: the approach to the fixed point is steep, not gentle. The bare parameters still vary as $1/a^\theta$ with $\theta \sim 2$. The cancellation of large numbers is still required. The fixed point explains *why* the cancellation happens (the trajectory demands it) but does not make the cancellation *smaller* (the exponent θ is not reduced).

The only way to eliminate the $1/a^2$ sensitivity is a symmetry that sets $\theta = 0$ (like chiral symmetry for fermion masses, or SUSY for the Higgs mass). Asymptotic safety does not provide this.

The hierarchy of concerns

From the lattice-first viewpoint, the concerns about quantum gravity fall into three levels, in order of decreasing physical relevance:

Level 1: bare-parameter sensitivity ($1/a^\theta$). A concern about the engine at its operating point. The Higgs bare mass requires a 30-digit cancellation at $a \sim \ell_P$. This is a structural fact about the engine’s calibration, not about a limit or an extrapolation. It is the most concrete concern, the most directly physical, and the only one that involves numbers the engine must actually get right.

If gravity is emergent: only m_H^2 has this problem ($\theta = 2$, no protection). G and Λ are Born-layer outputs and carry no bare-parameter sensitivity.

Level 2: economy of description (asymptotic safety). A concern about how many inputs the engine needs. If a fixed point exists: $n_{\text{rel}} \sim 3$ inputs suffice. If not: one input per coupling per scale. This affects the predictivity of the theory at $E \sim M_P$ but not the predictions at $E \ll M_P$ (where G and Λ suffice). A question about elegance more than physics.

Level 3: UV completion (the continuum limit). A concern about the rendering, not the engine. Does the $a \rightarrow 0$ limit exist as a Wightman QFT? This is a mathematical question about whether the screen image has an infinite-resolution limit. The engine runs at finite a ; the physics doesn’t depend on the answer.

	Concern	About	Physical?	Lattice-first status
1	$1/a^\theta$ sensitivity	Engine calibration	Yes (the engine is built this way)	Real for m_H^2 ; absent for G, Λ
2	Asymptotic safety	Number of inputs	Partially (predictivity at M_P)	Economy of description
3	UV completion	Continuum limit	No	A question about the rendering

The traditional literature often presents these concerns in the opposite order: UV completion first (“gravity is non- renormalisable!”), then asymptotic safety (“maybe a fixed point saves it!”), then the hierarchy problem (“but even then, why is m_H small?”). The lattice-first ordering reverses the priority: the bare-parameter sensitivity is the most concrete concern, the economy of description is secondary, and UV completion is a mathematical question with no physical consequence.

43.15 The one-way valve

The Fock-Born framework identifies $g_{\mu\nu}$ as a Born-layer output: the macroscopic, classical, maximally degenerate rendering of the graviton field. As a Born-layer output, $g_{\mu\nu}$ cannot be recycled as a Fock-layer input. This is the one-way valve (section ??): the engine produces $g_{\mu\nu}$; the screen displays it; neither the engine nor the screen can feed it back.

Every approach to quantum gravity that quantises $g_{\mu\nu}$ directly—canonical quantisation, the Wheeler–DeWitt equation, the path integral over metrics—recycles the Born-layer output into the Fock layer. This recycling is the origin of the problem of time (quantising the engine’s clock),

the problem of observables (what is a Born-layer observable when the metric is a Fock-layer variable?), and the non-renormalisability (the rendering-stability condition fails with infinitely many parameters).

This book does not propose a resolution. It observes that the one-way valve is violated by every existing approach to quantum gravity, and that the conceptual difficulties of quantum gravity (the problem of time, the information paradox, the landscape) may be symptoms of this violation rather than features of the correct theory.

The Fock layer operates on a flat lattice with spacing a . The Born layer renders $g_{\mu\nu}$ at $\lambda \gg a$. The lattice does not “live on” $g_{\mu\nu}$; it lives on the computational substrate. $g_{\mu\nu}$ is what the screen shows, not what the engine runs on. A theory of quantum gravity, in this framework, would be a theory of the Fock-layer engine (the lattice Hamiltonian $H(a)$ including the graviton as a spin-2 field on $\eta_{\mu\nu}$) whose Born-layer rendering produces the classical Einstein equation in the macroscopic limit—without ever quantising the rendered $g_{\mu\nu}$ itself.

44 Emergence of Collective Modes

44.1 Composite observables of fundamental fields

The Fock space $\mathcal{F}$ of a quantum field theory is built from creation operators $a_\alpha^\dagger$ of fundamental quanta (where α labels momentum, spin, flavour, colour, etc.). Every state in $\mathcal{F}$ is a superposition of definite-occupation states $|n_{\alpha_1}, n_{\alpha_2}, \dots\rangle$, and every observable is a polynomial of creation and annihilation operators.

Some observables are *elementary*: they create or destroy a single quantum ($a_{\mathbf{k}}^\dagger$, $a_{\mathbf{k}}$, or the number operator $\hat{n}_{\mathbf{k}} = a_{\mathbf{k}}^\dagger a_{\mathbf{k}}$). Their eigenvalues ($n = 0, 1, 2, \dots$) label the Fock basis directly, with no degeneracy: each eigenvalue determines a unique state (up to the occupation of other modes).

Other observables are *composite*: they involve products of multiple creation and annihilation operators at the same spacetime point. For a scalar field ϕ :

$$\hat{\Phi}(\mathbf{x}) = \hat{\phi}^2(\mathbf{x}) = \hat{\phi}^2(\mathbf{x}) - \langle 0 | \hat{\phi}^2 | 0 \rangle, \quad (44.1)$$

the normal-ordered square of the field. For a fermion ψ :

$$\hat{\Phi}(\mathbf{x}) = \hat{\psi}(\mathbf{x})\hat{\psi}(\mathbf{x}), \quad (44.2)$$

the normal-ordered scalar bilinear. In both cases: $\hat{\Phi}$ is a Hermitian operator on $\mathcal{F}$ with eigenvalues and eigensubspaces.

The degeneracy. The eigensubspace of $\hat{\Phi}$ at eigenvalue Φ_0 consists of all Fock states that produce the same value Φ_0 when $\hat{\Phi}$ is measured. Its dimension is the degeneracy:

$$d(\Phi_0) = \dim \mathcal{H}_{\Phi_0}. \quad (44.3)$$

For the elementary number operator $\hat{n}_{\mathbf{k}}$: the eigenvalue n specifies the occupation of mode $\mathbf{k}$, and the degeneracy is determined by the occupations of all other modes — which are completely unrestricted. The degeneracy is the dimension of the rest of the Fock space: enormous, but structured (the modes factorise).

For the composite $\hat{\Phi} = :\hat{\phi}^2:$: many different combinations of mode occupations can produce the same total ϕ^2 . The degeneracy counts these combinations. It has no simple factorisation: the composite entangles the mode occupations.

The microcanonical entropy.

$$S(\Phi_0) = k_B \ln d(\Phi_0). \quad (44.4)$$

This is the entropy of the composite: the number of bits of microscopic information hidden behind the macroscopic value Φ_0 . An elementary observable has zero compositional entropy (one microstate per eigenvalue, up to other modes). A composite observable has large entropy: many microstates per macrostate.

The Born-layer reading. When the Born rule renders an eigenvalue of $\hat{\Phi}$, it projects the Fock hidden state h onto the eigensubspace $\mathcal{H}_{\Phi_0}$ and normalises. The rendered value Φ_0 is the Born-layer output; the $d(\Phi_0)$ microstates compatible with Φ_0 are the Fock-layer ambiguity. The observer who measures only $\hat{\Phi}$ has *partial access*: they know Φ_0 but not which of the $d(\Phi_0)$ microstates the system is in. The entropy $S(\Phi_0)$ quantifies this ignorance.

44.2 The quasiparticle: a composite that propagates

The question of emergence is: can the composite $\hat{\Phi}$ behave as an independent particle? Does it have a definite energy-momentum relation, propagate coherently, scatter off other excitations, and be detected as a discrete quantum?

The answer is encoded in the *retarded Green's function* of the composite:

$$G_R(\mathbf{q}, t) = -i\theta(t) \langle 0 | [\hat{\Phi}_{\mathbf{q}}(t), \hat{\Phi}_{\mathbf{q}}^\dagger(0)] | 0 \rangle, \quad (44.5)$$

where $\hat{\Phi}_{\mathbf{q}} = \int d^{d-1}\mathbf{x} e^{-i\mathbf{q}\cdot\mathbf{x}} \hat{\Phi}(\mathbf{x})$ is the Fourier component at wavevector $\mathbf{q}$, and the expectation is in the full interacting vacuum $|0\rangle$ of H .

The retarded Green's function G_R is a Fock-layer quantity: a matrix element of two composite operators in the vacuum, evolved by the full Hamiltonian. It encodes the response of the system to a perturbation that creates one quantum of $\hat{\Phi}$ at $t = 0$ and asks: does the disturbance propagate?

The spectral function.

$$\rho(\omega, \mathbf{q}) = -2\text{Im} G_R(\omega + i0^+, \mathbf{q}). \quad (44.6)$$

The spectral function is the density of states of the composite at energy ω and momentum $\mathbf{q}$. It satisfies the sum rule $\int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \rho(\omega, \mathbf{q}) = 1$ (from the equal-time commutation relation of $\hat{\Phi}$).

The Lehmann representation. Inserting a complete set of energy eigenstates $|n\rangle$ of H :

$$G_R(\omega, \mathbf{q}) = \sum_n \frac{|\langle n | \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle|^2}{\omega - (E_n - E_0) + i\varepsilon} - \frac{|\langle n | \hat{\Phi}_{\mathbf{q}} | 0 \rangle|^2}{\omega + (E_n - E_0) + i\varepsilon}. \quad (44.7)$$

Each eigenstate $|n\rangle$ contributes a pole at $\omega = \pm(E_n - E_0)$ with residue $|\langle n | \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle|^2$: the matrix element of the composite between the vacuum and the excited state. The spectral function:

$$\rho(\omega, \mathbf{q}) = 2\pi \sum_n [|\langle n | \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle|^2 \delta(\omega - \omega_n) - |\langle n | \hat{\Phi}_{\mathbf{q}} | 0 \rangle|^2 \delta(\omega + \omega_n)], \quad (44.8)$$

where $\omega_n = E_n - E_0$. This is a sum of delta functions: one for each excited state that the composite can reach from the vacuum.

When is a quasiparticle present? If one of the states $|n\rangle$ has $\omega_n = \epsilon_{\mathbf{q}}$ (a definite energy-momentum relation) and a large matrix element $|\langle n | \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle|^2 \equiv Z \sim O(1)$, the spectral function has a sharp peak at $\omega = \epsilon_{\mathbf{q}}$ carrying a large fraction Z of the total spectral weight. This state IS the quasiparticle: it is the eigenstate of H that looks like “one quantum of $\hat{\Phi}$.”

In a many-body system: the exact delta functions in (44.8) merge into a continuous spectral function (the thermodynamic limit $L \rightarrow \infty$ turns the discrete sum into an integral). The quasiparticle peak becomes a Lorentzian of finite width Γ (because the state $|n\rangle$ is not an exact eigenstate in the continuum limit but a resonance embedded in the multi-particle continuum).

The Born-layer observable. The spectral function $\rho(\omega, \mathbf{q})$ is the most direct Born-layer observable of the composite: it is what an inelastic scattering experiment measures. In condensed matter: angle-resolved photoemission spectroscopy (ARPES) measures ρ for the electron; inelastic neutron scattering measures ρ for the phonon and magnon; Raman scattering measures ρ for optical phonons. In particle physics: the invariant-mass distribution $d\sigma/dM^2$ of the decay products of a resonance IS the spectral function in the relevant channel.

A sharp peak in $\rho \rightarrow$ the detector sees a particle. A broad continuum in $\rho \rightarrow$ the detector sees a statistical distribution. The spectral function bridges the Fock layer (the correlator G_R) to the Born layer (the detector response).

44.3 Worked example: the plasmon from the Lindhard function

We now compute the spectral function of a composite observable in a specific system — the electron density in a free Fermi gas — and show how interactions create a quasiparticle (the plasmon) out of a featureless continuum.

The composite. The density fluctuation at wavevector $\mathbf{q}$:

$$\hat{n}_{\mathbf{q}} = \sum_{\mathbf{k},s} c_{\mathbf{k}+\mathbf{q},s}^{\dagger} c_{\mathbf{k},s}, \quad (44.9)$$

where $c_{\mathbf{k},s}^{\dagger}$ creates an electron with momentum $\mathbf{k}$ and spin s . This is a composite of the fundamental electrons: it creates a particle-hole pair with total momentum $\mathbf{q}$.

The non-interacting response: the Lindhard function. For a free Fermi gas ($H = \sum_{\mathbf{k},s} \epsilon_{\mathbf{k}} c_{\mathbf{k},s}^{\dagger} c_{\mathbf{k},s}$, with $\epsilon_{\mathbf{k}} = \hbar^2 k^2 / (2m_e)$), the retarded density-density correlator:

$$\chi_0(\mathbf{q}, \omega) = \sum_{\mathbf{k},s} \frac{n_{\mathbf{k}} - n_{\mathbf{k}+\mathbf{q}}}{\omega - (\epsilon_{\mathbf{k}+\mathbf{q}} - \epsilon_{\mathbf{k}}) + i\varepsilon}, \quad (44.10)$$

where $n_{\mathbf{k}} = \theta(k_F - |\mathbf{k}|)$ is the Fermi function at $T = 0$. This is the *Lindhard function*: the response of the free Fermi gas to a density perturbation.

The spectral function of the free gas. $\text{Im} \chi_0(\mathbf{q}, \omega)$ is nonzero whenever the energy ω can excite a particle from below the Fermi surface ($|\mathbf{k}| < k_F$) to above it ($|\mathbf{k} + \mathbf{q}| > k_F$): the *particle-hole continuum*.

The continuum occupies the region $|\omega| < v_F q + q^2 / (2m_e)$ (for small q : $|\omega| < v_F q$, a wedge-shaped region in the (ω, q) plane, where $v_F = \hbar k_F / m_e$ is the Fermi velocity). Inside this region: $\rho_0(\omega, q) \neq 0$. Outside: $\rho_0 = 0$.

There is no sharp peak: ρ_0 is a smooth function across the continuum. No quasiparticle is present. The density fluctuation in the free gas is a continuum of particle-hole pairs, not a single propagating mode.

The degeneracy is enormous. Each point in the particle-hole continuum corresponds to many microscopic particle-hole configurations: different $\mathbf{k}$ values give the same $(\omega, \mathbf{q})$. The number of such configurations at fixed $(\omega, \mathbf{q})$ is the density of states of the continuum: $\rho_0(\omega, q) \propto \text{Im} \chi_0 \propto$ (number of ways to create a particle-hole pair with the given energy and momentum). The degeneracy is large; the spectral function is broad; the Born-layer rendering is featureless.

Turning on interactions: the RPA. Add the Coulomb interaction $V(\mathbf{q}) = 4\pi e^2 / q^2$ between electrons. The interacting density response (the random phase approximation, which resums the bubble diagrams to all orders):

$$\chi_{\text{RPA}}(\mathbf{q}, \omega) = \frac{\chi_0(\mathbf{q}, \omega)}{1 - V(\mathbf{q}) \chi_0(\mathbf{q}, \omega)}. \quad (44.11)$$

This is a geometric series: the density fluctuation creates a Coulomb potential, which creates another density fluctuation, which creates another potential, etc. The resummation is exact in the large- N limit (N = number of fermion species, here $N = 2$ for spin).

The interacting spectral function:

$$\rho_{\text{RPA}}(\omega, \mathbf{q}) = \frac{-2 \text{Im} \chi_0}{|1 - V \chi_0|^2}. \quad (44.12)$$

This has a pole when $1 - V(\mathbf{q})\chi_0(\mathbf{q}, \omega) = 0$: a zero of the dielectric function $\varepsilon(\mathbf{q}, \omega) = 1 - V\chi_0$.

At small q : the pole is at $\omega = \omega_p + O(q^2)$, where $\omega_p = \sqrt{4\pi n e^2 / m_e}$ is the plasma frequency. This is the *plasmon*: a collective oscillation of the electron density, with a definite energy-momentum relation $\omega_{\mathbf{q}} = \omega_p \sqrt{1 + 3v_F^2 q^2 / (5\omega_p^2)} + \dots$, a sharp spectral peak, and a large residue $Z \sim O(1)$.

The plasmon is a quasiparticle that does not exist in the free gas: it is *created by the interactions*. The Coulomb interaction reshuffles the spectral weight from the broad particle-hole continuum into a sharp peak above the continuum. The degeneracy of the continuum (many particle-hole configurations at each (ω, q)) is resolved into a single collective mode (the coherent density oscillation).

	Free gas ($V = 0$)	Interacting ($V \neq 0$)
Spectral function	Broad continuum ρ_0	Sharp peak at ω_p + continuum
Quasiparticle?	No	Yes: the plasmon
Degeneracy	Unresolved (many $\mathbf{k}$ per (ω, q))	Partially resolved (peak carries weight Z)
Born-layer output	Featureless absorption	Resonance peak in electron energy loss

The path integral computation. The same result on the Euclidean path-integral side: the density-density correlator is $\langle n_{\mathbf{q}}(\tau) n_{-\mathbf{q}}(0) \rangle_E = \chi_E(\mathbf{q}, \tau)$, and the RPA resummation is a geometric series of bubble diagrams:

$$\chi_{\text{RPA}}^E(q) = \frac{\chi_0^E(q)}{1 - V(q)\chi_0^E(q)}, \quad q = (\mathbf{q}, i\omega_n), \quad (44.13)$$

where $\omega_n = 2\pi n / \beta$ are the bosonic Matsubara frequencies and χ_0^E is the Euclidean Lindhard function. The physical spectral function is obtained by analytic continuation $i\omega_n \rightarrow \omega + i\varepsilon$.

The correspondence:

Path integral	Hamiltonian
$\chi_0^E(q)$ (Euclidean bubble)	$\chi_0(\mathbf{q}, \omega)$ (Lehmann sum)
$1 - V\chi_0^E = 0$ (pole)	Zero of $\varepsilon(\mathbf{q}, \omega)$
Analytic continuation $i\omega_n \rightarrow \omega$	Spectral decomposition
Same $\rho_{\text{RPA}}(\omega, \mathbf{q})$	Same $\rho_{\text{RPA}}(\omega, \mathbf{q})$

44.4 Worked example: the Cooper pair and the BCS condensate

The second example illustrates condensation: a composite that acquires a nonzero expectation value, reorganising the vacuum.

The composite. The pair operator at zero total momentum:

$$\hat{\Delta} = \sum_{\mathbf{k}} g_{\mathbf{k}} c_{-\mathbf{k}\downarrow} c_{\mathbf{k}\uparrow}, \quad (44.14)$$

where $g_{\mathbf{k}}$ is the pair wave function (e.g., $g_{\mathbf{k}} = 1$ for s -wave pairing near the Fermi surface). This is a composite of two fundamental electrons with opposite momenta and spins.

The pair operator creates a state with two fewer electrons. It is *not* a number-conserving observable: $\hat{\Delta}$ changes the particle number by -2 . Its expectation value $\langle \hat{\Delta} \rangle$ is zero in any state

with definite particle number (by superselection). But in the thermodynamic limit ($N \rightarrow \infty$): the BCS ground state has $\langle \hat{\Delta} \rangle \neq 0$ — the pair condenses.

The normal state ($T > T_c$). The pair susceptibility (the retarded Green's function of $\hat{\Delta}$):

$$\chi_{\text{pair}}(\omega) = \sum_{\mathbf{k}} \frac{1 - n_{\mathbf{k}} - n_{-\mathbf{k}}}{\omega - \epsilon_{\mathbf{k}} - \epsilon_{-\mathbf{k}} + i\varepsilon}. \quad (44.15)$$

At $T > T_c$: this has no pole. The spectral function is a continuum (the two-particle threshold at $\omega = 2\Delta_{\text{gap}}$ for the interacting case, or $\omega = 0$ for the non-interacting case). The pair is not a quasiparticle: it is a composite with a broad spectral distribution and no sharp peak.

The condensed state ($T < T_c$). Below T_c : an attractive interaction (phonon-mediated in conventional superconductors) makes the pairing susceptibility diverge at $\omega = 0$: the condition $1 + V_{\text{pair}}\chi_{\text{pair}}(0) = 0$ is satisfied. This is the *pairing instability*: the normal state is unstable to pair formation.

The ground state reorganises: the BCS vacuum

$$|\text{BCS}\rangle = \prod_{\mathbf{k}} (u_{\mathbf{k}} + v_{\mathbf{k}} c_{\mathbf{k}\uparrow}^\dagger c_{-\mathbf{k}\downarrow}^\dagger) |0\rangle, \quad (44.16)$$

with $|u_{\mathbf{k}}|^2 + |v_{\mathbf{k}}|^2 = 1$, is a superposition of states with different particle numbers. The condensate: $\langle \hat{\Delta} \rangle = \sum_{\mathbf{k}} u_{\mathbf{k}} v_{\mathbf{k}} \neq 0$.

Excitations above the condensate. Fluctuations of the pair field $\hat{\Delta}(x) = \Delta_0 + \delta\hat{\Delta}(x)$ around the condensate have two types:

The amplitude mode ($|\Delta|$ fluctuations): a massive mode with energy $2\Delta_0$ (the gap). This is the analogue of the Higgs mode: it costs energy to change the magnitude of the condensate.

The phase mode ($\arg \Delta$ fluctuations): a massless mode — the Goldstone boson of the broken $U(1)$ symmetry (particle number conservation). In a charged superconductor: the phase mode is “eaten” by the photon via the Anderson–Higgs mechanism, producing the Meissner effect ($m_\gamma \neq 0$ inside the superconductor). In a neutral superfluid (^4He , cold atoms): the phase mode is the Bogoliubov sound mode with $\omega = c_s q$.

Certification. The amplitude mode: $\Gamma/\epsilon \ll 1$ for $T \ll T_c$ (the mode is inside the gap, so it cannot decay into single-particle excitations). Certified. The Born-layer observable: a peak at $\omega = 2\Delta_0$ in the Raman scattering spectrum.

The phase mode (in a neutral superfluid): exactly massless, $Z = 1$ by Goldstone's theorem. Exactly certified (symmetry-protected). The Born-layer observable: the speed of sound c_s .

The degeneracy structure of the condensate. The BCS vacuum has a specific degeneracy structure. Define the pair number operator $\hat{N}_{\text{pair}} = \hat{\Delta}^\dagger \hat{\Delta}$. Before condensation ($T > T_c$): the vacuum has $\langle \hat{N}_{\text{pair}} \rangle \sim O(1)$, and the degeneracy at fixed pair number is that of the normal Fermi gas ($d \sim e^{S_{\text{Fermi}}/k_B}$).

After condensation ($T < T_c$): the vacuum has $\langle \hat{N}_{\text{pair}} \rangle \sim N$ (macroscopic), and the degeneracy is restructured. The pair occupation $v_{\mathbf{k}}^2$ is a smooth function of $\mathbf{k}$, and the entropy of the condensed state is lower than the normal state: $S_{\text{BCS}} < S_{\text{normal}}$ (the condensate resolves some of the microscopic degeneracy by ordering the pairs).

On the path-integral side: the BCS condensate is the saddle point Δ_0 of the effective action $S_{\text{eff}}[\Delta]$ (obtained by an HS transformation in the pair channel). The entropy reduction is visible in $V_{\text{eff}}(\Delta_0) < V_{\text{eff}}(0)$: the condensed state has lower free energy.

44.5 The effective propagator: direct computation

Both worked examples computed the composite's Green's function by resumming a series of diagrams (the bubble chain for the plasmon, the ladder series for the Cooper pair). We now state the general structure.

The path integral: diagram resummation

The Euclidean correlator of the composite $\hat{\Phi}$:

$$\Pi_E(q) = \int d^d x e^{iq \cdot x} \langle \hat{\Phi}(x) \hat{\Phi}(0) \rangle_E = \frac{\int \mathcal{D}\phi \Phi(x) \Phi(0) e^{-S_E}}{\int \mathcal{D}\phi e^{-S_E}}. \quad (44.17)$$

For $\hat{\Phi} = :\phi^2:$ in ϕ^4 theory: $\Pi_E(q) = \langle \phi^2(x) \phi^2(0) \rangle_{E, \text{conn}}$, the connected four-point function with two pairs of legs contracted. By Wick's theorem at lowest order: $\Pi_E^{(0)}(q) = 2[G_E(q)]^2$, where $G_E(q) = 1/(q^2 + m^2)$ is the free Euclidean propagator.

Including the interaction $\lambda\phi^4/4!$: the leading correction is the “bubble chain” — a geometric series of bubble diagrams:

$$\Pi_E(q) = \frac{\chi(q)}{1 - \tilde{\lambda}\chi(q)/2}, \quad \chi(q) = \int \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2 + m^2)((k+q)^2 + m^2)}, \quad (44.18)$$

where $\chi(q)$ is the single bubble (one-loop diagram) and the geometric series resums all iterations of the interaction. This is the RPA for the scalar field: the same structure as (44.11) for the electron gas, with the Coulomb potential $V(\mathbf{q})$ replaced by the coupling $\tilde{\lambda}/2$.

The pole: $\Pi_E(q)$ has a pole when $\tilde{\lambda}\chi(q_*)/2 = 1$. This is the certification condition on the path-integral side: the composite ϕ^2 has become a propagating mode.

The Hamiltonian: Feshbach projection within $\mathcal{F}$

On the Hamiltonian side, the composite's effective propagator is computed by Feshbach projection *within the original Fock space* $\mathcal{F}_\phi$ (no enlargement, no auxiliary field):

Define the projector P onto the “composite sector”: states of the form $\hat{\Phi}_{\mathbf{q}}^\dagger |0\rangle$ (one quantum of the composite above the vacuum). For $\hat{\Phi} = :\phi^2:$ these are two-particle states $a_{\mathbf{k}}^\dagger a_{\mathbf{q}-\mathbf{k}}^\dagger |0\rangle$, summed over $\mathbf{k}$ with appropriate coefficients.

Define $Q = 1 - P$: the rest of $\mathcal{F}$ (four-particle states, six-particle states, vacuum contributions, etc.).

The effective Green's function of the composite:

$$G_{\text{eff}}(\mathbf{q}, \omega) = \langle 0 | \hat{\Phi}_{\mathbf{q}} \frac{1}{\omega - H + E_0 + i\varepsilon} \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle. \quad (44.19)$$

Using the Feshbach decomposition $H = \begin{pmatrix} H_{PP} & V \\ V^\dagger & H_{QQ} \end{pmatrix}$:

$$G_{\text{eff}}(\mathbf{q}, \omega) = \frac{1}{\omega - \epsilon_{\mathbf{q}}^{(0)} - \Sigma(\mathbf{q}, \omega)}, \quad (44.20)$$

where $\epsilon_{\mathbf{q}}^{(0)} = \langle 0 | \hat{\Phi}_{\mathbf{q}} H_{PP} \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle$ is the bare energy of the composite (the energy of a non-interacting pair with total momentum $\mathbf{q}$), and the self-energy:

$$\Sigma(\mathbf{q}, \omega) = \langle 0 | \hat{\Phi}_{\mathbf{q}} V \frac{1}{\omega - H_{QQ} + E_0} V^\dagger \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle. \quad (44.21)$$

The self-energy encodes how the composite couples to the multi-particle continuum (the $\mathcal{Q}$ sector): it emits additional quanta into $\mathcal{Q}$, they propagate, and are reabsorbed.

The real part of Σ shifts the energy (mass correction). The imaginary part gives the width (decay rate into the continuum):

$$\Gamma(\mathbf{q}, \omega) = -2 \text{Im} \Sigma(\mathbf{q}, \omega) = 2\pi \sum_{n \in \mathcal{Q}} |\langle n | V^\dagger \hat{\Phi}_{\mathbf{q}}^\dagger | 0 \rangle|^2 \delta(\omega - \omega_n). \quad (44.22)$$

This is Fermi's golden rule: the decay rate into the multi-particle states of $\mathcal{Q}$.

The correspondence

The path-integral and Hamiltonian computations give the same effective propagator:

Object	Path integral	Hamiltonian
The correlator	$\Pi_E(q) = \langle \Phi(x) \Phi(0) \rangle_E$	$G_{\text{eff}}(\mathbf{q}, \omega)$ via Lehmann
The self-energy	Irreducible bubble $\Sigma_{\text{irr}}(q)$	$\Sigma = V(E - H_{\mathcal{Q}\mathcal{Q}})^{-1} V^\dagger$
The pole condition	$1 - \tilde{\lambda} \chi(q)/2 = 0$	$\omega - \epsilon^{(0)} - \Sigma(\omega) = 0$
The width	$\text{Im} \chi(q_{\text{pole}})$	Fermi's golden rule (44.22)
The residue Z	$[1 - \tilde{\lambda} \chi'(q)/2]^{-1}$	$[1 - \partial_\omega \Sigma]^{-1}$
Analytic continuation	$i\omega_n \rightarrow \omega + i\varepsilon$	Already in real frequency
Born-layer output	$\rho(\omega) = -2 \text{Im} \Pi_E _{i\omega_n \rightarrow \omega}$	$\rho(\omega) = -2 \text{Im} G_{\text{eff}}$

Both methods compute the same spectral function $\rho(\omega, \mathbf{q})$: the Born-layer observable that encodes whether the composite is a quasiparticle. The path integral works in Euclidean signature (the bubble integral is convergent) and requires analytic continuation to get the real-frequency spectral function. The Hamiltonian works directly in real frequency but requires specifying the projector P and the coupling V . Same physics, two languages.

44.6 Certification: the quantitative criteria

We can now state the certification conditions as consequences of the effective propagator, derived in both languages.

The effective propagator in the general case:

$$G_{\text{eff}}(\omega, \mathbf{q}) = \frac{Z_{\mathbf{q}}}{\omega - \epsilon_{\mathbf{q}} + i\Gamma_{\mathbf{q}}/2} + G_{\text{inc}}(\omega, \mathbf{q}), \quad (44.23)$$

where $\epsilon_{\mathbf{q}}$ is the renormalised energy, $\Gamma_{\mathbf{q}}$ is the width, $Z_{\mathbf{q}}$ is the residue (quasiparticle weight), and G_{inc} is the incoherent continuum.

The three certification numbers.

$N_{\text{osc}} = \epsilon_{\mathbf{q}}/\Gamma_{\mathbf{q}}$: *oscillations before decay*. The quasiparticle oscillates at frequency ϵ and decays at rate Γ . If $N_{\text{osc}} \gg 1$: the mode completes many cycles before decaying — it is a well-defined particle. If $N_{\text{osc}} \sim 1$: it decays within one cycle — not a particle.

$Z_{\mathbf{q}}$: *the spectral weight in the peak*. The fraction of the total spectral weight carried by the quasiparticle peak. If $Z \sim 1$: the peak exhausts the sum rule — nearly all the composite's response is in the quasiparticle. If $Z \rightarrow 0$: the weight has leaked into the continuum — the quasiparticle has dissolved.

$\partial_\omega \text{Im } \Sigma|_\epsilon$: the self-energy's frequency dependence. If $\text{Im } \Sigma \propto \omega^2$ near the pole (Fermi liquid): $\Gamma \propto \omega^2 \ll \omega$ at low energies — certification improves as $\omega \rightarrow 0$. If $\text{Im } \Sigma \propto \omega$ (marginal Fermi liquid): $\Gamma \propto \omega$ at all energies — certification never improves.

Condition	Criterion	Physical meaning	Failure mode
Sharp pole	$\Gamma \ll \epsilon$	$N_{\text{osc}} \gg 1$	Strange metal: $\Gamma \sim \epsilon$
Large residue	$Z \sim O(1)$	Peak dominates spectral weight	Luttinger liquid: $Z \rightarrow 0$
Factorisation	$\text{Im } \Sigma \propto \omega^n, n \geq 2$	Width vanishes at low ω	Marginal FL: $n = 1$

Examples ranked by certification quality.

Mode	Γ/ϵ	Z	N_{osc}	Status
Photon	0	1	∞	Exactly certified (gauge)
Pion $\pi^\pm$	10^{-9}	~ 1	10^9	Sharply certified (chiral + weak)
Plasmon	~ 0.01	~ 0.8	~ 100	Well certified (RPA)
$\rho(770)$	0.19	~ 0.5	5	Marginal (broad resonance)
$\sigma/f_0(500)$	~ 1	~ 0.2	~ 1	Poorly certified
Strange metal e	ω/ω	$\rightarrow 0$	1	Not certified

44.7 Entropy and the Born-layer measure of certification

The spectral function $\rho(\omega)$ has two components: the quasiparticle peak (weight Z) and the incoherent continuum (weight $1 - Z$). These carry different Born-layer meanings.

The peak: resolved degeneracy. The quasiparticle peak at $\omega = \epsilon_{\mathbf{q}}$ represents a single collective excitation. An observer who detects the quasiparticle (e.g., by inelastic scattering at the resonance energy) resolves the composite into one quantum number: the occupation of the collective mode. The entropy associated with the peak is small: $S_{\text{peak}} \sim k_B \ln(\text{few})$.

The continuum: unresolved degeneracy. The incoherent continuum represents the multi-particle states that the composite can reach: the $d(\Phi_0)$ microstates compatible with the macroscopic value Φ_0 . An observer who cannot distinguish the quasiparticle from the continuum sees the full degeneracy: $S_{\text{cont}} \sim k_B \ln d_{\text{micro}}$.

Certification as degeneracy resolution.

	Well-certified ($\Gamma \ll \epsilon, Z \sim 1$)	Failed certification ($\Gamma \sim \epsilon, Z \rightarrow 0$)
Spectral fn	Sharp peak + weak continuum	Broad continuum, no peak
Observer sees	A particle (one quantum number)	A thermal bath (many microstates)
Entropy	$S \approx S_{\text{peak}}$ (small)	$S \approx k_B \ln d_{\text{micro}}$ (large)
Temperature	Low (few accessible states)	Emerges: $T = (\partial S / \partial E)^{-1}$
Degeneracy	Resolved (peak isolates the mode)	Unresolved (continuum masks the mode)

Certification is the resolution of degeneracy. The Born-to-Fock promotion replaces e^{S/k_B}

degenerate microstates by a single quantum number. When it works ($Z \sim 1$): the particle picture holds. When it fails ($Z \rightarrow 0$): the observer is left with the full degeneracy, the entropy is large, and temperature emerges from the unresolved microscopic degrees of freedom.

44.8 The vacuum state and condensation

When $\langle 0|\hat{\Phi}|0\rangle \neq 0$: the vacuum has a nonzero expectation value of the composite. This is *condensation*: the macroscopic occupation of a single quantum state by the composite degrees of freedom.

The condensate as a saddle point. On the path-integral side: the condensate Φ_0 is the field value that dominates the functional integral. The effective potential $V_{\text{eff}}(\Phi_0) = -\frac{1}{V\beta} \ln \int \mathcal{D}\phi \delta(\Phi[\phi] - \Phi_0) e^{-S_E}$ has a minimum at $\Phi_0 \neq 0$: the saddle-point configuration has a nonzero composite.

On the Hamiltonian side: the ground state $|\Omega\rangle$ of H has $\langle \Omega|\hat{\Phi}|\Omega\rangle = \Phi_0 \neq 0$: the composite operator has a nonzero vacuum expectation value.

The two descriptions agree: saddle point of $V_{\text{eff}} = \text{VEV of } \hat{\Phi} \text{ in } |\Omega\rangle$.

Symmetry breaking. If the Hamiltonian has a symmetry G under which $\hat{\Phi}$ transforms non-trivially, and $\langle \hat{\Phi} \rangle \neq 0$ breaks G to a subgroup H : the condensate spontaneously breaks $G \rightarrow H$.

Goldstone's theorem: for each broken continuous generator ($\dim G/H$ generators), there is a massless mode — a Goldstone boson. These are the phase fluctuations of the condensate: the modes that rotate Φ_0 within the degenerate vacuum manifold G/H .

The amplitude mode: fluctuations of $|\Phi|$ around Φ_0 . Massive, with mass $\sim \sqrt{V''_{\text{eff}}(\Phi_0)}$.

The degeneracy structure of the condensed vacuum. Before condensation: the vacuum has $\langle \hat{\Phi} \rangle = 0$, and the degeneracy at $\Phi = 0$ is maximal (all configurations contribute).

After condensation: the vacuum selects $\Phi_0 \neq 0$, and the degeneracy is restructured. The ground-state manifold is G/H (the set of equivalent condensate orientations). The entropy of the condensed state is lower: the condensate resolves some of the microscopic degeneracy by imposing long-range order.

The Born-layer observables of the condensate. The condensate Φ_0 is a Born-layer observable: it is measurable (e.g., the superfluid density in ^4He , the gap Δ_0 in a superconductor, the chiral condensate $\langle \bar{\psi}\psi \rangle$ in QCD via the GMOR relation). The Goldstone modes are Born-layer observables: their dispersion relation $\omega(q)$ is measured by inelastic scattering. The amplitude mode (if certified) is a Born-layer observable: a peak in the spectral function at $\omega \sim 2\Phi_0$.

The condensate is the most dramatic Born-to-Fock promotion: the vacuum itself reorganises around the composite, promoting it from a fluctuating Born-layer observable to the defining feature of the Fock-layer ground state.

Act I establishes what certification is, how it is computed (path integral and Hamiltonian), what it means thermodynamically (degeneracy resolution), and what happens when the composite condenses. Act II introduces the Hubbard–Stratonovich transformation as the systematic tool for implementing the Born-to-Fock promotion, making all the above computations more efficient.

Act I computed the composite's Green's function directly: resumming bubble diagrams (path integral) or Feshbach projection within $\mathcal{F}$ (Hamiltonian). Both required identifying the composite

channel by hand and resumming an infinite series. The HS transformation automates this: it introduces the composite as an independent field, integrates out the fundamentals exactly, and delivers the effective action in closed form.

44.9 The HS transformation for scalars: Euclidean path integral

The microscopic theory. A real scalar ϕ with quartic self-interaction, in Euclidean signature:

$$S_E[\phi] = \int d^d x \left[\frac{1}{2}(\partial\phi)^2 + \frac{m^2}{2}\phi^2 + \frac{\lambda}{4!}\phi^4 \right], \quad Z_E = \int \mathcal{D}\phi e^{-S_E}. \quad (44.24)$$

Step 1: the identity. Define $\tilde{\lambda} = \lambda/3$. The Gaussian integral identity:

$$\boxed{e^{-\frac{\tilde{\lambda}}{8} \int (\phi^2)^2} = \int \mathcal{D}\sigma e^{-\frac{1}{2\tilde{\lambda}} \int \sigma^2 + \frac{1}{2} \int \sigma \phi^2}}. \quad (44.25)$$

The quartic ϕ^4 is replaced by a linear coupling $\sigma\phi^2/2$ plus a Gaussian weight $\sigma^2/(2\tilde{\lambda})$ for the auxiliary field $\sigma(x)$.

Step 2: the modified partition function.

$$Z_E = \int \mathcal{D}\sigma \mathcal{D}\phi e^{-\int [\frac{1}{2}\phi(-\nabla^2 + m^2 + \sigma)\phi + \frac{\sigma^2}{2\tilde{\lambda}}]}. \quad (44.26)$$

At fixed σ : the ϕ action is quadratic with shifted mass $m^2 \rightarrow m^2 + \sigma(x)$.

Step 3: integrate out ϕ . The Gaussian integral over ϕ :

$$\int \mathcal{D}\phi e^{-\frac{1}{2} \int \phi D \phi} = (\det D)^{-1/2}, \quad D = -\nabla^2 + m^2 + \sigma. \quad (44.27)$$

The effective action for σ :

$$\boxed{S_{\text{eff}}^E[\sigma] = \frac{1}{2\tilde{\lambda}} \int \sigma^2 + \frac{1}{2} \text{Tr} \ln(-\nabla^2 + m^2 + \sigma)}. \quad (44.28)$$

First term: the Gaussian prior (the “cost” of the auxiliary field). Second term: the scalar determinant (the ϕ fluctuations in the background σ).

Step 4: the on-shell identification. The equation of motion $\delta S_{\text{eff}}/\delta\sigma = 0$:

$$\frac{\sigma(x)}{\tilde{\lambda}} = -\frac{1}{2} \langle x | (-\nabla^2 + m^2 + \sigma)^{-1} | x \rangle = -\frac{1}{2} \langle \phi^2(x) \rangle_\sigma. \quad (44.29)$$

On shell: $\sigma = -\frac{\tilde{\lambda}}{2} \langle \phi^2 \rangle$. The auxiliary field IS the composite ϕ^2 , up to a coupling constant. This is why the HS transformation is the Born-to-Fock promotion: it takes the Born-layer observable ϕ^2 and gives it its own field variable σ .

Step 5: the gap equation. For a homogeneous saddle point $\sigma(x) = \sigma_0$:

$$\frac{\sigma_0}{\tilde{\lambda}} = -\frac{1}{2} \int \frac{d^d k}{(2\pi)^d} \frac{1}{k^2 + M^2}, \quad M^2 = m^2 + \sigma_0. \quad (44.30)$$

This self-consistency equation determines σ_0 (and hence the effective mass M) as a function of the coupling $\tilde{\lambda}$ and the cutoff.

Step 6: fluctuations and the propagator. Expand $\sigma = \sigma_0 + \delta\sigma$. The quadratic action:

$$S^{(2)}[\delta\sigma] = \frac{1}{2} \int \frac{d^d q}{(2\pi)^d} \delta\sigma(-q) \Pi^{-1}(q) \delta\sigma(q), \quad (44.31)$$

$$\Pi^{-1}(q) = \frac{1}{\tilde{\lambda}} + \frac{1}{2}\chi(q), \quad \chi(q) = \int \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2 + M^2)((k+q)^2 + M^2)}. \quad (44.32)$$

This is the same $\Pi^{-1}(q)$ as the direct calculation (44.18) of Act I: the HS transformation automatically resums the bubble chain. The infinite geometric series that required manual resummation in §44.5 is now a single Gaussian integral.

44.10 The Gaussian noise of the auxiliary field

In the Euclidean path integral, $\sigma(x)$ fluctuates around σ_0 with weight $e^{-S_{\text{eff}}^E[\sigma]}$. The two-point correlator:

$$\langle \delta\sigma(x) \delta\sigma(y) \rangle_E = \Pi(x-y) = \int \frac{d^d q}{(2\pi)^d} \Pi(q) e^{iq \cdot (x-y)}. \quad (44.33)$$

This is the *Gaussian noise* of the auxiliary field.

The noise IS the composite's fluctuation. Since $\sigma = -\frac{\tilde{\lambda}}{2}\phi^2$ on shell:

$$\langle \delta\sigma(x) \delta\sigma(y) \rangle_E = \frac{\tilde{\lambda}^2}{4} \langle \phi^2(x) \phi^2(y) \rangle_{E, \text{conn}}. \quad (44.34)$$

The Gaussian noise of σ in the Euclidean path integral equals the connected correlator of the composite ϕ^2 in the microscopic theory. The HS transformation does not create fluctuations; it reorganises the existing fluctuations of ϕ^2 into an independent field.

The noise spectrum and certification. The Fourier transform $\Pi(q)$ determines the noise power at each momentum. If $\Pi(q)$ has a pole at $q^2 = -m_\sigma^2$ (Euclidean): the noise has a resonant enhancement at the pole — the quasiparticle peak. The pole in the noise spectrum IS the quasiparticle: it is the mode at which the composite fluctuates most strongly.

If $\Pi(q)$ has no pole: the noise is featureless (no resonance). The composite fluctuates with a broad, structureless spectrum. No quasiparticle.

44.11 Block diagonalisation for scalars: Hamiltonian

The Hamiltonian counterpart introduces σ as an operator on an enlarged Fock space, then projects out the microscopic degrees of freedom.

The enlarged Fock space. $\mathcal{F}_{\text{enl}} = \mathcal{F}_\phi \otimes \mathcal{F}_\sigma$, with independent creation operators $a_{\mathbf{k}}^\dagger$ (for ϕ) and $b_{\mathbf{q}}^\dagger$ (for σ).

The enlarged Hamiltonian.

$$H_{\text{enl}} = H_\phi^{(\text{free})} + \frac{1}{2\tilde{\lambda}} \int d^{d-1}\mathbf{x} \hat{\sigma}^2 + \frac{1}{2} \int d^{d-1}\mathbf{x} \hat{\sigma} : \hat{\phi}^2 :. \quad (44.35)$$

The quartic ϕ^4 is replaced by the cubic vertex $\hat{\sigma} : \hat{\phi}^2 : / 2$. The bare σ sector has only a mass term $\hat{\sigma}^2 / (2\tilde{\lambda})$: no kinetic energy. The bare σ does not propagate.

The block decomposition.

$\mathcal{P}$: states with one σ -quantum and the ϕ sector in its vacuum.

$\mathcal{Q}$: all other states (ϕ -pairs, multi-particle configurations, no σ).

$$H_{\text{enl}} = \begin{pmatrix} H_{PP} & V \\ V^\dagger & H_{QQ} \end{pmatrix}, \quad V = \frac{1}{2} \int \hat{\sigma} : \hat{\phi}^2 :. \quad (44.36)$$

The Feshbach effective Hamiltonian.

$$H_{\text{eff}}^{(\sigma)}(E) = H_{PP} + V \frac{1}{E - H_{QQ}} V^\dagger. \quad (44.37)$$

The self-energy in the single- σ sector:

$$\Sigma_\sigma(\mathbf{q}, E) = \frac{1}{4} \sum_{\mathbf{k}} \frac{|\langle \phi_{\mathbf{k}} \phi_{\mathbf{q}-\mathbf{k}} | : \hat{\phi}^2 : | 0 \rangle|^2}{E - \omega_{\mathbf{k}} - \omega_{\mathbf{q}-\mathbf{k}}}. \quad (44.38)$$

The σ quantum virtually creates a ϕ -pair (via the vertex V), the pair propagates in $\mathcal{Q}$ (via $1/(E - H_{QQ})$), and is reabsorbed (via $V^\dagger$).

The induced kinetic term. The bare σ in H_{PP} has mass but no kinetic energy. Expand $\Sigma_\sigma(\mathbf{q}, E)$ in powers of $\mathbf{q}$:

$$\Sigma_\sigma(\mathbf{q}, E) = \Sigma_\sigma(0, E) + \frac{1}{2} Z_\sigma q^2 + \dots \quad (44.39)$$

The q^2 term is the induced kinetic energy: the ϕ -pair cloud dresses the bare σ with a gradient energy. Propagation is entirely induced by the microscopic fluctuations. This is the Hamiltonian counterpart of the kinetic term that emerges from expanding $\text{Tr} \ln(-\nabla^2 + m^2 + \sigma)$ in the path integral.

44.12 The correspondence: Euclidean noise and Hamiltonian vacuum

The Euclidean path integral and the Hamiltonian describe the same physics in two languages. For the auxiliary field σ , the correspondence is:

The Gaussian noise = the vacuum fluctuation.

Path integral: $\sigma(x)$ fluctuates in the Euclidean functional integral with correlator $\langle \delta\sigma \delta\sigma \rangle_E = \Pi(x - y)$. This is the noise of a classical-looking stochastic field, weighted by $e^{-S_{\text{eff}}}$.

Hamiltonian: the vacuum $|\Omega\rangle$ of H_{enl} is an entangled state in $\mathcal{F}_\phi \otimes \mathcal{F}_\sigma$:

$$|\Omega\rangle = \sum_n c_n |\phi_n\rangle \otimes |\sigma_n\rangle. \quad (44.40)$$

The vacuum fluctuation of $\hat{\sigma}$: $\langle \Omega | (\hat{\sigma} - \sigma_0)^2 | \Omega \rangle = \sum_n |c_n|^2 (\sigma_n - \sigma_0)^2$.

The two are related by Wick rotation:

$$\langle \delta\sigma(x) \delta\sigma(y) \rangle_E \big|_{\tau=it} = \langle \Omega | T \{ \hat{\sigma}(x) \hat{\sigma}(y) \} | \Omega \rangle. \quad (44.41)$$

The Euclidean noise, analytically continued to real time, is the time-ordered vacuum correlator.

The saddle point = the VEV.

Path integral: σ_0 minimises S_{eff}^E (the most probable value under $e^{-S_{\text{eff}}}$).

Hamiltonian: $\sigma_0 = \langle \Omega | \hat{\sigma} | \Omega \rangle$ (the ground-state expectation value).

Same number: the saddle point of the Euclidean weight = the VEV of the operator.

The vacuum phase. $\sigma_0 = 0$: the symmetric phase. The vacuum preserves any symmetry under which $\hat{\Phi}$ transforms.

$\sigma_0 \neq 0$: the broken phase. The vacuum has a nonzero composite: $\langle \Omega | : \hat{\phi}^2 : | \Omega \rangle \neq 0$ (scalar) or $\langle \Omega | \hat{\psi} \hat{\psi} | \Omega \rangle \neq 0$ (fermion).

On the path-integral side: the symmetric/broken phases correspond to the single-well/double-well structure of $V_{\text{eff}}(\sigma)$. On the Hamiltonian side: they correspond to the symmetric/ symmetry-broken ground state of H_{eff} .

	Path integral (Euclidean)	Hamiltonian
Auxiliary field	$\sigma(x)$: integration variable	$\hat{\sigma}$: operator on $\mathcal{F}_\sigma$
Noise / fluctuation	$\langle \delta\sigma\delta\sigma \rangle_E = \Pi(x-y)$	$\langle \Omega (\hat{\sigma} - \sigma_0)^2 \Omega \rangle$
Saddle point / VEV	$\min S_{\text{eff}}^E$	$\langle \Omega \hat{\sigma} \Omega \rangle$
Phase (symmetric)	Single-well V_{eff}	Symmetric $ \Omega\rangle$
Phase (broken)	Double-well V_{eff}	$ \Omega\rangle$ breaks symmetry
Effective action	$S_{\text{eff}}^E[\sigma]$	$H_{\text{eff}}^{(\sigma)}(E)$
Certification	Pole in $\Pi(q)$	Gap in spectrum of H_{eff}

44.13 The HS transformation for fermions

The fermionic case (the Nambu–Jona-Lasinio model) follows the same structure with three important differences.

The microscopic theory.

$$S_E[\psi] = \int d^d x [\bar{\psi}(\not{\partial} + m)\psi + \frac{g}{2}(\bar{\psi}\psi)^2]. \quad (44.42)$$

The HS identity.

$$e^{-\frac{g}{2} \int (\bar{\psi}\psi)^2} = \int \mathcal{D}\sigma e^{-\frac{1}{2g} \int \sigma^2 + \int \sigma \bar{\psi}\psi}. \quad (44.43)$$

Integrating out ψ . The action is quadratic in ψ at fixed σ . The Grassmann integral:

$$\int \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{-\int \bar{\psi}(\not{\partial} + m + \sigma)\psi} = \det(\not{\partial} + m + \sigma). \quad (44.44)$$

The effective action:

$$S_{\text{eff}}^{E,(\psi)}[\sigma] = \frac{1}{2g} \int \sigma^2 - \text{Tr} \ln(\not{\partial} + m + \sigma). \quad (44.45)$$

The three differences from the scalar case.

1. *The sign of the determinant.* Scalar: $+\frac{1}{2} \text{Tr} \ln D$ (from $(\det D)^{-1/2}$, the inverse square root for a real boson). Fermion: $-\text{Tr} \ln D$ (from $\det D$, the determinant for a complex fermion with the Grassmann sign). The minus sign makes the fermion-loop contribution *opposite* in sign to the boson-loop contribution, which is the origin of the boson-fermion cancellation in supersymmetric theories.

2. *Chiral symmetry breaking.* For $m = 0$: the Lagrangian (44.42) has a chiral symmetry $\psi \rightarrow e^{i\alpha\gamma_5}\psi$. The gap equation for the fermion case:

$$\frac{\sigma_0}{g} = N_c N_f \int \frac{d^d k}{(2\pi)^d} \frac{m + \sigma_0}{k^2 + (m + \sigma_0)^2}. \quad (44.46)$$

For sufficiently large g : $\sigma_0 \neq 0$ even at $m = 0$. The fermion bilinear condenses: $\langle \bar{\psi}\psi \rangle = \sigma_0/g \neq 0$. The fermion acquires a dynamical mass $M = m + \sigma_0$.

3. *Goldstone bosons.* If chiral symmetry is continuous ($SU(2)_L \times SU(2)_R$ for $N_f = 2$), the condensate breaks it to $SU(2)_V$, producing three massless pions — the Goldstone bosons of chiral symmetry breaking. These appear as poles in the pseudoscalar channel of $S_{\text{eff}}[\sigma, \vec{\pi}]$ (where $\vec{\pi}$ are the pseudoscalar auxiliary fields introduced by an HS transformation in the $\bar{\psi}i\gamma_5\vec{\tau}\psi$ channel).

44.14 Block diagonalisation for fermions: Hamiltonian

The enlarged Fock space. $\mathcal{F}_{\text{enl}} = \mathcal{F}_\psi \otimes \mathcal{F}_\sigma$, with the Hamiltonian:

$$H_{\text{enl}} = H_\psi^{(\text{free})} + \frac{1}{2g} \int \hat{\sigma}^2 + \int \hat{\sigma} : \hat{\psi}\hat{\psi} :. \quad (44.47)$$

The block decomposition. $\mathcal{P}$: one σ -quantum, fermion sector in the Dirac sea. $\mathcal{Q}$: fermion-antifermion pairs, no σ .

The Feshbach self-energy.

$$\Sigma_\sigma(\mathbf{q}, E) = \sum_{\mathbf{k}, s, s'} \frac{|\langle \bar{\psi}_{\mathbf{k}, s} \psi_{\mathbf{q}-\mathbf{k}, s'} | : \hat{\psi}\hat{\psi} : | 0 \rangle|^2}{E - E_{\mathbf{k}} - E_{\mathbf{q}-\mathbf{k}}}, \quad (44.48)$$

where $E_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + M^2}$ ($M = m + \sigma_0$). The fermion-loop sign (-1) enters through the anticommutation of the creation operators in the matrix element, making the contribution attractive.

The effective potential. The Dirac sea energy as a function of σ_0 :

$$V_{\text{eff}}(\sigma_0) = \frac{\sigma_0^2}{2g} - N_c N_f \int \frac{d^{d-1}\mathbf{k}}{(2\pi)^{d-1}} \sqrt{|\mathbf{k}|^2 + (m + \sigma_0)^2}. \quad (44.49)$$

The gap equation $\partial V_{\text{eff}}/\partial \sigma_0 = 0$ reproduces (44.46). The Goldstone bosons are the modes with $m_\pi = 0$ (at $m = 0$): the poles in the pseudoscalar channel of H_{eff} .

44.15 The two cases compared

	Scalar (ϕ^4)	Fermion (NJL)
Composite	$\hat{\sigma} \propto : \hat{\phi}^2 :$	$\hat{\sigma} \propto : \hat{\psi}\hat{\psi} :$
HS vertex	$\hat{\sigma} : \hat{\phi}^2 :$ (cubic)	$\hat{\sigma} : \hat{\psi}\hat{\psi} :$ (Yukawa)
Determinant	$+\frac{1}{2} \text{Tr} \ln D$	$-\text{Tr} \ln D$
Loop sign	$+1$ (boson)	-1 (fermion: attractive)
Condensation	Mass shift $m^2 \rightarrow m^2 + \sigma_0$	Chiral breaking $M = m + \sigma_0$
Goldstone boson	None (no continuous sym. broken)	Pion (0^{-+} , from chiral)
Gap equation	$\sigma_0/\tilde{\lambda} = -\frac{1}{2} \langle \phi^2 \rangle$	$\sigma_0/g = N_c N_f \langle \bar{\psi}\psi \rangle$
Decay channel	$\sigma \rightarrow \phi\phi$	$\sigma \rightarrow \bar{\psi}\psi$

Same architecture: enlarge $\rightarrow$ block-decompose $\rightarrow$ project $\rightarrow$ certify.

44.16 The fate of the auxiliary field

The effective action $S_{\text{eff}}[\sigma]$ (path integral) or $H_{\text{eff}}(\sigma)$ (Hamiltonian) determines whether σ becomes a certified particle.

Case 1: propagating pole ($Z > 0$, $\Gamma \ll m_\sigma$). The determinant generates a kinetic term $\frac{1}{2}Z_\sigma(\partial\sigma)^2$ and a finite mass m_σ^2 : the propagator has a sharp pole. The auxiliary field is a particle. Examples: the Higgs boson, the phonon, the magnon.

On the Born layer: the detector sees a resonance peak at $\omega = m_\sigma$. The degeneracy of the composite is resolved: the quasiparticle peak carries weight $Z \sim 1$.

Case 2: broad or absent ($Z \rightarrow 0$, $\Gamma \sim m_\sigma$). No sharp pole. The auxiliary field is a mathematical convenience, not a particle. The HS transformation is exact (the identity (44.25) always holds), but the promotion to a particle is not physical.

On the Born layer: the detector sees a broad continuum, no resonance. The degeneracy is unresolved.

Case 3: condensation ($\sigma_0 \neq 0$). The effective potential develops a minimum at $\sigma_0 \neq 0$: the composite condenses. Fluctuations around σ_0 may or may not be certified (a condensate can exist without a sharp amplitude mode).

On the Born layer: the condensate itself is observable (the BCS gap, the chiral condensate $\langle \bar{\psi}\psi \rangle$, the superfluid density). The Goldstone modes (if any) are exactly certified ($\Gamma = 0$, $Z = 1$ by Goldstone's theorem).

Case	Pole?	Certified?	Born-layer signature
Propagating	Sharp ($\Gamma \ll m$)	Yes	Resonance peak in $\rho(\omega)$
Broad	None ($Z \rightarrow 0$)	No	Featureless continuum
Condensate mode	+ Sharp around σ_0	Yes	Gap + coherence peak
Condensate only	$\sigma_0 \neq 0$, no mode	Partial	Order parameter, no excitation
Goldstone	Exact at $\omega = c_s q$	Exact	Sound mode / pion

Act II has shown that the HS transformation automates the Born-to-Fock promotion: introduce the composite as a field, integrate out the fundamentals, and read off the effective action. The transformation is exact; the certification is physical. Act III now asks: what happens at finite temperature?

44.17 The origin of temperature: partial access

Temperature does not appear in the Fock-Born architecture. The Schrödinger equation $\dot{h}_t = -iHh_t$ has no temperature; the Born rule $P(\lambda) = \|P_\lambda h\|^2$ has no temperature. Temperature arises when the Born-layer observer has partial access to the Fock-layer engine.

The derivation. Consider a system with $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$ in a pure state $|\Psi\rangle$. The observer measures only observables of subsystem A . The reduced density matrix:

$$\hat{\rho}_A = \text{Tr}_B |\Psi\rangle \langle \Psi|. \quad (44.50)$$

The partial trace sums over all states of B : it is the Born-layer consequence of measuring an observable with degeneracy $\dim \mathcal{H}_B$. The full state $|\Psi\rangle$ is pure (entropy zero); the reduced state $\hat{\rho}_A$ is mixed (entropy $S_A = -k_B \text{Tr}(\hat{\rho}_A \ln \hat{\rho}_A) \geq 0$).

For a chaotic many-body system in an energy eigenstate $|E\rangle$ (the eigenstate thermalisation hypothesis, ETH):

$$\hat{\rho}_A \approx \frac{e^{-\beta H_A}}{Z_A(\beta)}, \quad Z_A(\beta) = \text{Tr}_A e^{-\beta H_A}, \quad (44.51)$$

where $\beta = 1/(k_B T)$ is determined by the energy: $\langle H \rangle_E = -\partial \ln Z_A / \partial \beta$. The subsystem looks thermal because the environment B acts as a heat bath, and the partial trace averages over the bath's microstates.

Temperature is a Born-layer concept. The Fock layer has one vector, evolving unitarily. It has no temperature. Temperature arises in the Born layer when the observer measures a composite observable with degeneracy $d \sim e^{S/k_B}$: the rendered statistics are thermal, with $\beta = \partial S / \partial E$ controlling the degeneracy gradient.

This connects directly to certification (section 44.7): when a composite is well-certified ($Z \sim 1$), the observer resolves the quasiparticle quantum number and sees low entropy. When certification fails ($Z \rightarrow 0$), the observer sees the full degeneracy, the entropy is large, and temperature emerges.

44.18 The Euclidean periodic time

The thermal partition function $Z(\beta) = \text{Tr} e^{-\beta H}$ has a path-integral representation that connects temperature to geometry.

The path integral.

$$Z(\beta) = \int_{\phi(\mathbf{x},0)=\phi(\mathbf{x},\beta)} \mathcal{D}\phi e^{-S_E[\phi]}, \quad (44.52)$$

where $S_E = \int_0^\beta d\tau \int d^{d-1}\mathbf{x} \mathcal{L}_E$ and the field is periodic: $\phi(\mathbf{x}, \tau + \beta) = \phi(\mathbf{x}, \tau)$ (bosons) or antiperiodic: $\psi(\mathbf{x}, \tau + \beta) = -\psi(\mathbf{x}, \tau)$ (fermions).

The Euclidean time τ is compactified to a circle S_β^1 of circumference β . The spacetime is $S_\beta^1 \times \mathbb{R}^{d-1}$: a cylinder.

The Matsubara frequencies. Periodicity restricts the Fourier modes:

$$\phi(\mathbf{x}, \tau) = \frac{1}{\beta} \sum_n \phi_n(\mathbf{x}) e^{i\omega_n \tau}, \quad \omega_n = \begin{cases} 2\pi n / \beta & \text{bosons} \\ (2n + 1)\pi / \beta & \text{fermions.} \end{cases} \quad (44.53)$$

At $T \rightarrow 0$ ($\beta \rightarrow \infty$): S_β^1 decompactifies, $\frac{1}{\beta} \sum_n \rightarrow \int \frac{d\omega}{2\pi}$, and the zero-temperature formalism is recovered.

The unification. The two origins of temperature are the same statement:

Algebraic: partial trace over B produces $\hat{\rho}_A = e^{-\beta H_A} / Z_A$. The trace requires summing over all states of B .

Geometric: the trace $\text{Tr} e^{-\beta H}$ is implemented by periodic boundary conditions $\phi(\beta) = \phi(0)$. The period is $\beta = 1/(k_B T)$.

The algebraic partial trace (summing over unmeasured degrees of freedom) IS the geometric periodicity (compactifying the Euclidean time direction). Temperature is the circumference of the thermal circle.

44.19 The HS transformation at finite temperature

Apply the HS transformation inside the thermal partition function (44.52). For the fermion case:

$$Z(\beta) = \int \mathcal{D}\sigma \, e^{-S_{\text{eff}}^E[\sigma; \beta]}, \quad S_{\text{eff}}^E[\sigma; \beta] = \frac{1}{2g} \int_0^\beta d\tau \int d^{d-1}\mathbf{x} \, \sigma^2 - \text{Tr} \ln D_E[\sigma; \beta], \quad (44.54)$$

where $D_E = \not{\partial}_E + m + \sigma$ is the Euclidean Dirac operator and the trace includes the Matsubara sum.

For a homogeneous saddle point σ_0 : the effective potential $V_{\text{eff}}(\sigma_0; \beta) = S_{\text{eff}}^E[\sigma_0; \beta]/(\beta V)$.

Evaluating the Matsubara sum. The fermion contribution:

$$\frac{1}{\beta V} \text{Tr} \ln D_E = N_c N_f \frac{1}{\beta} \sum_n \int \frac{d^{d-1}\mathbf{k}}{(2\pi)^{d-1}} \ln(\omega_n^2 + E_{\mathbf{k}}^2), \quad (44.55)$$

where $\omega_n = (2n+1)\pi/\beta$ and $E_{\mathbf{k}} = \sqrt{|\mathbf{k}|^2 + (m + \sigma_0)^2}$.

Using $\frac{1}{\beta} \sum_n \ln(\omega_n^2 + E^2) = E + \frac{2}{\beta} \ln(1 + e^{-\beta E}) + \text{const}$:

$$V_{\text{eff}}(\sigma_0; \beta) = \frac{\sigma_0^2}{2g} - N_c N_f \int \frac{d^{d-1}\mathbf{k}}{(2\pi)^{d-1}} \left[E_{\mathbf{k}} + \frac{2}{\beta} \ln(1 + e^{-\beta E_{\mathbf{k}}}) \right]. \quad (44.56)$$

First bracket: the vacuum energy (the Dirac sea, $T = 0$, UV-divergent, regulated by the lattice).
Second bracket: the thermal correction ($T > 0$, UV-finite, the Fermi-Dirac thermal occupation).

44.20 The gap equation and equation of state

The gap equation at finite T . $\partial V_{\text{eff}}/\partial \sigma_0 = 0$:

$$\frac{\sigma_0}{g} = N_c N_f \int \frac{d^{d-1}\mathbf{k}}{(2\pi)^{d-1}} \frac{m + \sigma_0}{E_{\mathbf{k}}} [1 - 2n_F(\beta E_{\mathbf{k}})], \quad (44.57)$$

where $n_F(x) = 1/(e^x + 1)$.

At $T = 0$ ($n_F \rightarrow 0$): the gap equation reduces to the zero-temperature version (44.46). At $T > 0$: the thermal factor $[1 - 2n_F] < 1$ reduces the right-hand side, weakening the condensate. At $T = T_c$: $\sigma_0(T_c) = 0$ — the chiral symmetry is restored.

For the scalar case: the gap equation has $n_B(\beta E)$ instead of n_F , and the thermal factor is $[1 + 2n_B] > 1$ (the Bose enhancement strengthens the effect at high T).

The thermodynamic potential. The grand potential per unit volume (generalised to include a chemical potential μ):

$$\Omega(T, \mu) = V_{\text{eff}}(\sigma_0(T, \mu); T, \mu), \quad (44.58)$$

evaluated at the self-consistent $\sigma_0(T, \mu)$.

The equation of state.

$$P = -\Omega \quad (\text{pressure}), \quad (44.59)$$

$$n = -\frac{\partial \Omega}{\partial \mu} = N_c N_f \int \frac{d^{d-1}\mathbf{k}}{(2\pi)^{d-1}} [n_F(\beta(E_{\mathbf{k}} - \mu)) - n_F(\beta(E_{\mathbf{k}} + \mu))] \quad (\text{number density}), \quad (44.60)$$

$$\epsilon = -P + Ts + \mu n \quad (\text{energy density}), \quad (44.61)$$

$$s = -\frac{\partial \Omega}{\partial T} \quad (\text{entropy density}). \quad (44.62)$$

At high T ($T \gg \sigma_0, m$): the condensate melts ($\sigma_0 \rightarrow 0$), the fermions are effectively massless, and the equation of state approaches the Stefan–Boltzmann limit:

$$P_{\text{SB}} = N_c N_f \frac{7\pi^2}{180} T^4 \quad (\text{massless fermions}). \quad (44.63)$$

The Born-layer observables. Every line above is a Born-layer observable: pressure (what a manometer reports), density (what a counter reports), energy density (what a calorimeter reports), entropy density (what a thermometer and calorimeter together report). The gap equation is a Fock-layer self-consistency condition; the equation of state is its Born-layer output.

The Hamiltonian side gives the same equation of state: the partition function $Z(\beta) = \text{Tr } e^{-\beta H}$ is computed from the spectrum of $H_{\text{eff}}^{(\sigma)}$, and the thermodynamic potential $\Omega = -T \ln Z/V$ is the same function of T and μ .

44.21 Thermodynamics versus hydrodynamics: the Euclidean/Minkowski choice

The Euclidean effective action $S_{\text{eff}}^E[\sigma; \beta]$ at finite temperature contains two classes of Born-layer information, accessed by different analytic continuations.

Stay Euclidean: equilibrium thermodynamics. The Matsubara correlator $G_E(\tau, \mathbf{q}) = \langle \sigma(\tau, \mathbf{q}) \sigma(0, -\mathbf{q}) \rangle_E$ gives equilibrium quantities directly: the effective potential, the gap equation, the phase diagram (T_c vs. μ), the equation of state. No analytic continuation needed.

Rotate to Minkowski: real-time dynamics. The retarded Green’s function $G_R(\omega, \mathbf{q})$ is obtained by analytic continuation $i\omega_n \rightarrow \omega + i\varepsilon$:

$$G_E(i\omega_n, \mathbf{q})|_{i\omega_n \rightarrow \omega + i\varepsilon} = G_R(\omega, \mathbf{q}). \quad (44.64)$$

The spectral function $\rho(\omega, \mathbf{q}) = -2 \text{Im } G_R$ encodes the real-time dynamics: the collective-mode dispersion, the damping rates, and the transport coefficients.

Transport coefficients via Kubo formulae. The shear viscosity η , bulk viscosity ζ , and conductivity σ_{el} :

$$\eta = \lim_{\omega \rightarrow 0} \frac{\rho_{T_{12}T_{12}}(\omega, \mathbf{0})}{2\omega}, \quad \zeta = \frac{1}{9} \lim_{\omega \rightarrow 0} \frac{\rho_{T_{ii}T_{jj}}(\omega, \mathbf{0})}{2\omega}, \quad (44.65)$$

where ρ_{AB} is the spectral function of the stress-tensor components.

The comparison.

	Stay Euclidean	Rotate to Minkowski
Signature	$S_\beta^1 \times \mathbb{R}^{d-1}$	$\mathbb{R}^{1,d-1}$ at temperature T
Computes	$Z(\beta), F, P(T), s(T), T_c$	$G_R(\omega), \rho(\omega), \eta, \zeta, c_s$
Method	Monte Carlo (direct sampling)	Analytic continuation (indirect)
Physics	Equilibrium thermodynamics	Near-equilibrium hydrodynamics
Born-layer output	What a thermometer reports	What a viscometer reports

Both are Born-layer observables computed from the same Fock-layer engine. The Euclidean formulation gives equilibrium directly (Monte Carlo samples $e^{-S_E} \geq 0$). The Minkowski formulation gives dynamics but requires inverting an ill-posed integral (the analytic continuation from discrete Matsubara frequencies to continuous real frequencies).

The Hamiltonian side has no continuation problem. On the Hamiltonian side: the retarded Green's function $G_R(\omega)$ is computed directly from the spectrum of H , without analytic continuation. The spectral function is $\rho(\omega) = 2\pi \sum_n |\langle n|\hat{\Phi}|0\rangle|^2 \delta(\omega - \omega_n)$: it is the Lehmann representation, already in real frequency. Transport coefficients follow from ρ via the Kubo formulae.

This is another instance of the two-language correspondence:

Observable	Path integral	Hamiltonian
Equilibrium (P, s, T_c)	$S_{\text{eff}}^E[\sigma; \beta]$ directly	$Z = \text{Tr } e^{-\beta H_{\text{eff}}}$
Transport (η, ζ)	Analytic continuation of G_E	Lehmann spectral function directly
Gap equation	$\partial V_{\text{eff}}/\partial \sigma_0 = 0$	$\partial E_{\text{vac}}/\partial \sigma_0 = 0$
Phase transition	Change of saddle-point structure	Level crossing in H_{eff}

Acts I–III have established the complete framework for emergence: how composite observables of fundamental fields acquire particle-like properties (Act I), how the HS transformation and block diagonalisation systematise the Born-to-Fock promotion (Act II), and how temperature, entropy, and the equation of state arise from the degeneracy of the composite at finite temperature (Act III). The examples (phonon, pion, skyrmion, universality, the description hierarchy) follow in Part E.

45 String Theory and Loop Quantum Gravity

Two programmes attempt to quantise gravity. String theory replaces point particles with one-dimensional strings and derives gravity as a consistency condition of the worldsheet. Loop quantum gravity quantises the spatial geometry directly, using the Holst action as the starting point. This section develops both systematically, then examines each from the Fock-Born framework.

Part I: String Theory

45.1 From hadrons to gravitons: the historical arc

String theory did not begin as a theory of quantum gravity.

1968–70: dual models for hadrons. Veneziano (1968) wrote an amplitude for meson scattering with the correct Regge behaviour (s -channel and t -channel poles dual to each other). Nambu, Nielsen, and Susskind (1970) realised that Veneziano’s formula described the scattering of one-dimensional strings: the hadronic flux tube between a quark and an antiquark.

1974: the graviton. The closed string spectrum always contains a massless spin-2 state. Scherk and Schwarz (1974) proposed reinterpreting this state as the *graviton*: the quantum of the gravitational field. The string tension $T = 1/(2\pi\alpha')$, originally $\sim \Lambda_{\text{QCD}}^2$ for hadrons, was promoted to $\sim M_{\text{P}}^2$ for gravity. String theory became a candidate theory of quantum gravity.

1984: the Green–Schwarz breakthrough. Green and Schwarz showed that type I superstring theory in $d = 10$ with gauge group $SO(32)$ is free of gauge and gravitational anomalies. This anomaly cancellation—a precise algebraic miracle—convinced the community that string theory is a consistent quantum theory of gravity. The “first superstring revolution” followed.

45.2 The worldsheet action

A string sweeps out a two-dimensional worldsheet Σ in a D -dimensional spacetime. The Polyakov action:

$$S_P = -\frac{1}{4\pi\alpha'} \int_{\Sigma} d^2\sigma \sqrt{-h} h^{ab} \partial_a X^\mu \partial_b X^\nu G_{\mu\nu}(X), \quad (45.1)$$

where $X^\mu(\sigma^0, \sigma^1)$ are the spacetime coordinates of the string, h_{ab} is the worldsheet metric (an auxiliary field), and $G_{\mu\nu}$ is the background spacetime metric.

The Polyakov action has three local worldsheet symmetries: diffeomorphisms ($\sigma^a \rightarrow \sigma'^a$) and Weyl rescaling ($h_{ab} \rightarrow e^{2\omega} h_{ab}$). In conformal gauge $h_{ab} = e^{2\phi} \eta_{ab}$, on a flat background $G_{\mu\nu} = \eta_{\mu\nu}$:

$$S = -\frac{1}{4\pi\alpha'} \int d^2\sigma \eta^{ab} \partial_a X^\mu \partial_b X_\mu : \quad (45.2)$$

D free massless scalars on a 2D Minkowski worldsheet.

For strings propagating on a general background, the action includes couplings to the Kalb–Ramond field $B_{\mu\nu}$ and the dilaton Φ :

$$S = -\frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{-h} [h^{ab} G_{\mu\nu}(X) \partial_a X^\mu \partial_b X^\nu + \varepsilon^{ab} B_{\mu\nu}(X) \partial_a X^\mu \partial_b X^\nu + \alpha' R^{(2)} \Phi(X)], \quad (45.3)$$

where ε^{ab} is the worldsheet Levi-Civita symbol and $R^{(2)}$ is the worldsheet Ricci scalar. The three massless background fields ($G_{\mu\nu}, B_{\mu\nu}, \Phi$) are the universal sector of every string theory.

45.3 Conformal field theory on the worldsheet

The gauge-fixed worldsheet theory is a two-dimensional conformal field theory (CFT). In two dimensions, the conformal group is infinite-dimensional, making the CFT structure extremely constraining: the symmetry is powerful enough to determine correlation functions, operator spectra, and consistency conditions that have no analogue in higher dimensions.

The Virasoro algebra

Introduce the complex worldsheet coordinate $z = e^{\tau_E + i\sigma}$ (where τ_E is the Euclidean worldsheet time and $\sigma \in [0, 2\pi)$ parametrises the string). The energy-momentum tensor of the free bosonic fields $X^\mu(z, \bar{z})$:

$$T(z) = -\frac{1}{\alpha'} : \partial X^\mu(z) \partial X_\mu(z) :, \quad \bar{T}(\bar{z}) = -\frac{1}{\alpha'} : \bar{\partial} X^\mu(\bar{z}) \bar{\partial} X_\mu(\bar{z}) :, \quad (45.4)$$

where $\partial = \partial/\partial z$, $\bar{\partial} = \partial/\partial \bar{z}$, and the colons denote normal ordering.

Expand $T(z)$ in Laurent modes:

$$T(z) = \sum_{n \in \mathbb{Z}} L_n z^{-n-2}, \quad L_n = \oint \frac{dz}{2\pi i} z^{n+1} T(z). \quad (45.5)$$

The modes L_n satisfy the *Virasoro algebra*:

$$[L_m, L_n] = (m - n) L_{m+n} + \frac{c}{12} m(m^2 - 1) \delta_{m+n,0}, \quad (45.6)$$

where c is the *central charge*. The first term is the classical conformal algebra (the Witt algebra); the second term is the *conformal anomaly* — a quantum correction that cannot be removed by any redefinition of the generators.

For D free bosons: $c = D$ (each X^μ contributes $c = 1$). The central charge counts the effective number of degrees of freedom of the CFT.

Primary fields and conformal dimension. A field $\mathcal{O}(z, \bar{z})$ is a *conformal primary* of weight $(h, \bar{h})$ if it transforms under conformal maps $z \rightarrow f(z)$ as:

$$\mathcal{O}(z, \bar{z}) \rightarrow (\partial f)^h (\bar{\partial} \bar{f})^{\bar{h}} \mathcal{O}(f(z), \bar{f}(\bar{z})). \quad (45.7)$$

Equivalently: the OPE with the energy-momentum tensor has the standard form $T(z)\mathcal{O}(w) \sim h\mathcal{O}(w)/(z-w)^2 + \partial\mathcal{O}(w)/(z-w)$.

The conformal weights $(h, \bar{h})$ determine the spacetime quantum numbers:

Mass: $m^2 = \frac{2}{\alpha'}(h + \bar{h} - a - \bar{a})$, where $a, \bar{a}$ are the normal-ordering constants.

Spin: $s = |h - \bar{h}|$, the angular momentum on the worldsheet, which becomes the spacetime spin.

The operator product expansion

The OPE of the free bosons:

$$\partial X^\mu(z) \partial X^\nu(w) \sim -\frac{\alpha'}{2} \frac{\eta^{\mu\nu}}{(z-w)^2} + \text{regular}. \quad (45.8)$$

The singular part ($\sim 1/(z-w)^2$) encodes the short-distance behaviour; the regular part contains no singularity as $z \rightarrow w$. The OPE replaces the notion of “multiplication at a point” in the 2D

CFT: the product of two operators at nearby points is expanded as a sum of local operators with singular coefficients.

The OPE is the most important computational tool in 2D CFT. It determines:

- (i) *Commutation relations*: the commutator of two mode operators $[L_m, L_n]$ is extracted from the singular part of the $T(z)T(w)$ OPE.
- (ii) *Correlation functions*: the n -point function is computed by successive OPEs, reducing to lower-point functions.
- (iii) *Scattering amplitudes*: the string scattering amplitude is the correlation function of vertex operators, computed by OPEs.

The state-operator correspondence

Every state $|\psi\rangle$ in the worldsheet Hilbert space corresponds to a local operator $\mathcal{O}_\psi(z)$ on the worldsheet, and vice versa:

$$|\psi\rangle = \lim_{z \rightarrow 0} \mathcal{O}_\psi(z) |0\rangle. \quad (45.9)$$

The operator $\mathcal{O}_\psi$, inserted at $z = 0$ (the origin of the complex plane, corresponding to the infinite past on the cylinder), creates the state $|\psi\rangle$ from the $SL(2, \mathbb{C})$ vacuum $|0\rangle$.

This correspondence is exact (not an approximation) and is specific to 2D CFT: it relies on the conformal map from the cylinder ($\tau_E \in (-\infty, \infty)$, $\sigma \in [0, 2\pi)$) to the complex plane ($z = e^{\tau_E + i\sigma}$), which maps the infinite past $\tau_E \rightarrow -\infty$ to the origin $z = 0$.

Examples:

$|0; p\rangle$ (the tachyon with momentum p) $\leftrightarrow e^{ip \cdot X(0)}$ (the tachyon vertex operator).

$\alpha_{-1}^\mu |0; p\rangle$ (a massless vector) $\leftrightarrow \partial X^\mu(0) e^{ip \cdot X(0)}$.

$\alpha_{-1}^\mu \bar{\alpha}_{-1}^\nu |0; p\rangle$ (the graviton) $\leftrightarrow \partial X^\mu(0) \bar{\partial} X^\nu(0) e^{ip \cdot X(0)}$.

45.4 Vertex operators

Vertex operators are the bridge between the worldsheet CFT and spacetime physics: each spacetime particle corresponds to a specific conformal primary on the worldsheet.

The tachyon vertex operator.

$$V_{\text{tachyon}}(z, \bar{z}) = e^{ip \cdot X(z, \bar{z})}. \quad (45.10)$$

The exponential $e^{ip \cdot X}$ is a conformal primary of weight $(h, \bar{h}) = (\alpha' p^2/4, \alpha' p^2/4)$. The condition that V be a $(1, 1)$ primary (required for the integrated vertex operator $\int d^2z V$ to be conformally invariant) gives:

$$\alpha' p^2/4 = 1 \quad \implies \quad m^2 = p^2 = -\frac{4}{\alpha'} \quad (\text{tachyonic: } m^2 < 0). \quad (45.11)$$

The tachyon signals an instability of the bosonic string vacuum (the vacuum decays). It is absent in the superstring (removed by the GSO projection).

The graviton vertex operator.

$$V_{\text{graviton}}(z, \bar{z}) = \epsilon_{\mu\nu} : \partial X^\mu(z) \bar{\partial} X^\nu(\bar{z}) e^{ip \cdot X(z, \bar{z})} :, \quad (45.12)$$

where $\epsilon_{\mu\nu}$ is the polarisation tensor (symmetric, traceless: $\epsilon_{\mu\nu} = \epsilon_{\nu\mu}$, $\eta^{\mu\nu} \epsilon_{\mu\nu} = 0$).

The conformal weight: $h = 1 + \alpha' p^2/4$, $\bar{h} = 1 + \alpha' p^2/4$. The $(1, 1)$ condition gives $p^2 = 0$: the graviton is massless.

The transversality condition $p^\mu \epsilon_{\mu\nu} = 0$ follows from the requirement that V be BRST-closed (physical state condition): $Q_B V = 0$ requires $p^\mu \epsilon_{\mu\nu} = 0$. This is the string-theoretic origin of the gauge invariance $\epsilon_{\mu\nu} \rightarrow \epsilon_{\mu\nu} + p_\mu \xi_\nu + p_\nu \xi_\mu$ (linearised diffeomorphisms).

The Kalb–Ramond vertex operator.

$$V_B(z, \bar{z}) = B_{\mu\nu} : \partial X^\mu(z) \bar{\partial} X^\nu(\bar{z}) e^{ip \cdot X(z, \bar{z})} :, \quad (45.13)$$

with $B_{\mu\nu} = -B_{\nu\mu}$ (antisymmetric). Same conformal weight as the graviton: massless, $p^2 = 0$.

The dilaton vertex operator.

$$V_\Phi(z, \bar{z}) = (\eta_{\mu\nu} - p_\mu \bar{p}_\nu - \bar{p}_\mu p_\nu) : \partial X^\mu(z) \bar{\partial} X^\nu(\bar{z}) e^{ip \cdot X(z, \bar{z})} :, \quad (45.14)$$

where $\bar{p}^\mu$ is an auxiliary null vector. The dilaton is the trace part of the $\partial X \bar{\partial} X$ vertex operator: it is the scalar component of the closed-string massless sector.

Higher-level vertex operators. At oscillator level $N = \bar{N} = 2$: the vertex operators involve $\partial^2 X$, $(\partial X)^2$, and higher products. These create massive states with $m^2 = 2/\alpha'$ (first massive level), $m^2 = 4/\alpha'$ (second level), etc. The number of states at each level grows exponentially (Hagedorn growth): $d_N \sim N^{-a} e^{c\sqrt{N}}$.

The vertex operator determines the scattering. The n -point scattering amplitude is the worldsheet correlation function of n vertex operators:

$$\mathcal{A}_n = g_s^{n-2} \int \prod_{i=4}^n d^2 z_i \left\langle \prod_{j=1}^n V_j(z_j, \bar{z}_j) \right\rangle_{S^2}, \quad (45.15)$$

where the first three vertex operators are fixed by the $SL(2, \mathbb{C})$ symmetry of the sphere (the conformal Killing group), and the remaining $n-3$ are integrated over the sphere. The correlation function is computed by Wick contractions using the OPE (45.8).

45.5 The critical dimension

The critical dimension $D = 26$ (bosonic) or $D = 10$ (superstring) is not a postulate but a *consistency condition*: Weyl invariance at the quantum level requires the total central charge to vanish.

The Weyl anomaly

Classical Weyl invariance ($h_{ab} \rightarrow e^{2\omega} h_{ab}$ leaves the Polyakov action invariant) is broken at the quantum level by the conformal anomaly. The trace of the worldsheet energy-momentum tensor, which vanishes classically ($T^a_a = 0$), acquires a quantum correction:

$$\langle T^a_a \rangle = \frac{c_{\text{total}}}{12} R_{\text{ws}}, \quad (45.16)$$

where R_{ws} is the worldsheet Ricci scalar and c_{total} is the total central charge of all worldsheet fields, including the reparametrisation ghosts.

If $c_{\text{total}} \neq 0$: the Weyl anomaly is nonzero, the gauge symmetry is broken, and the worldsheet metric h_{ab} acquires a dynamical degree of freedom (the Liouville mode). The resulting theory is Liouville gravity, not a consistent string background.

If $c_{\text{total}} = 0$: the Weyl anomaly cancels, the gauge symmetry is preserved, and the worldsheet metric can be fully gauge-fixed. This is the critical string.

The ghost system

Gauge-fixing the worldsheet reparametrisation invariance ($\xi^a(\sigma) \rightarrow \xi^a(\sigma) + \epsilon^a(\sigma)$) by the Faddeev–Popov procedure introduces ghost fields $b_{ab}(z)$ (a symmetric traceless tensor of weight $(2, 0)$) and $c^a(z)$ (a vector of weight $(-1, 0)$):

$$S_{\text{ghost}} = \frac{1}{2\pi} \int d^2z (b_{zz} \partial_{\bar{z}} c^z + b_{\bar{z}\bar{z}} \partial_z c^{\bar{z}}). \quad (45.17)$$

The ghost system is a free CFT with central charge:

$$c_{\text{ghost}} = -26 \quad (\text{from } b, c \text{ of weights } 2, -1). \quad (45.18)$$

The value -26 is determined by the conformal weights of the ghosts: $c_{b,c} = -3(2\lambda - 1)^2 + 1$ for ghosts of weight $(\lambda, 1 - \lambda)$; with $\lambda = 2$: $c = -26$.

The bosonic string: $D = 26$

Total central charge:

$$c_{\text{total}} = \underbrace{D}_{\text{matter } X^\mu} + \underbrace{(-26)}_{\text{ghosts } b,c} = D - 26. \quad (45.19)$$

Weyl invariance $c_{\text{total}} = 0$ requires:

$$\boxed{D = 26 \quad (\text{bosonic string critical dimension}).} \quad (45.20)$$

Alternative derivation: the light-cone. In light-cone quantisation: the physical degrees of freedom are the $D - 2$ transverse oscillators α_n^i ($i = 1, \dots, D - 2$). The Lorentz algebra $[M^{i-}, M^{j-}]$ closes only if $D = 26$ and the normal-ordering constant $a = 1$. The calculation uses:

$$\sum_{n=1}^{\infty} n = \zeta(-1) = -\frac{1}{12} \quad (\text{zeta regularisation}), \quad (45.21)$$

giving $a = \frac{D-2}{24} \cdot 1 = 1$ at $D = 26$. This is the same zeta regularisation that produces the Lüscher term $-\pi(D - 2)/(24R)$ for the QCD flux tube (section 41.8).

The superstring: $D = 10$

Adding worldsheet supersymmetry: each X^μ has a fermionic partner ψ^μ (a Majorana–Weyl fermion on the worldsheet).

Matter central charge: $c_{\text{matter}} = D + D/2 = 3D/2$ (D from X^μ , $D/2$ from ψ^μ : each ψ contributes $c = 1/2$).

Ghost central charge: $c_{\text{ghost}} = -26 + 11 = -15$ (-26 from b, c as before, $+11$ from the superconformal ghosts β, γ of weights $(3/2, -1/2)$: $c_{\beta,\gamma} = +11$).

Total:

$$c_{\text{total}} = \frac{3D}{2} - 15. \quad (45.22)$$

Weyl invariance requires:

$$\boxed{D = 10 \quad (\text{superstring critical dimension}).} \quad (45.23)$$

In light-cone: $D - 2 = 8$ transverse oscillators, $a = 1/2$ (superstring), and the super-Poincaré algebra closes at $D = 10$.

45.6 The string spectrum and the graviton

Mode expansion and oscillators

Expand the worldsheet fields in modes. For the closed bosonic string (with $\sigma \in [0, 2\pi)$, periodic boundary conditions):

$$X^\mu(\tau, \sigma) = x^\mu + \alpha' p^\mu \tau + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} (\alpha_n^\mu e^{-in(\tau-\sigma)} + \bar{\alpha}_n^\mu e^{-in(\tau+\sigma)}). \quad (45.24)$$

The left-movers α_n^μ and right-movers $\bar{\alpha}_n^\mu$ are independent oscillator sets satisfying:

$$[\alpha_m^\mu, \alpha_n^\nu] = m \eta^{\mu\nu} \delta_{m+n,0}, \quad [\bar{\alpha}_m^\mu, \bar{\alpha}_n^\nu] = m \eta^{\mu\nu} \delta_{m+n,0}. \quad (45.25)$$

For $n > 0$: α_n^μ is an annihilation operator ($\alpha_n^\mu |0\rangle = 0$); α_{-n}^μ is a creation operator. The vacuum $|0; p\rangle$ has centre-of-mass momentum p^μ .

The left-moving and right-moving number operators:

$$N = \sum_{n=1}^{\infty} \alpha_{-n} \cdot \alpha_n, \quad \bar{N} = \sum_{n=1}^{\infty} \bar{\alpha}_{-n} \cdot \bar{\alpha}_n. \quad (45.26)$$

The mass-shell condition

The Virasoro constraints $L_0 = \bar{L}_0 = 0$ (the worldsheet Hamiltonian annihilates physical states) give:

$$\boxed{\alpha' m^2 = 2(N - a) = 2(\bar{N} - \bar{a}), \quad N = \bar{N} \text{ (level matching)},} \quad (45.27)$$

where $a = \bar{a} = 1$ (bosonic) or $a = \bar{a} = 1/2$ (superstring) is the normal-ordering constant (from $\sum_{n=1}^{\infty} n = \zeta(-1) = -1/12$, with $D - 2$ transverse oscillators: $a = (D - 2)/24 = 24/24 = 1$ at $D = 26$).

Level matching ($N = \bar{N}$): a closed string must have equal numbers of left-moving and right-moving excitations (from the periodicity of σ).

The bosonic string spectrum

Level $N = \bar{N} = 0$ (tachyon). $m^2 = -4/\alpha' < 0$: a tachyonic scalar. Signals vacuum instability. Removed in the superstring by the GSO projection.

Level $N = \bar{N} = 1$ (massless states). The states $\alpha_{-1}^\mu \bar{\alpha}_{-1}^\nu |0; p\rangle$ have $m^2 = 0$. The tensor $\alpha_{-1}^\mu \bar{\alpha}_{-1}^\nu$ decomposes under the Lorentz group into three irreducible representations:

Symmetric traceless ($G_{\mu\nu}$): $(D - 2)(D - 1)/2 - 1$ states. This is a massless spin-2 field: the **graviton**. Its vertex operator is (45.12). The existence of a massless spin-2 state in every closed string theory is the string-theoretic origin of gravity.

Antisymmetric ($B_{\mu\nu}$): $(D - 2)(D - 3)/2$ states. The **Kalb–Ramond** 2-form. It couples to the string worldsheet: $\int_{\Sigma} B_{\mu\nu} dX^\mu \wedge dX^\nu$.

Trace (Φ): 1 state. The **dilaton**. Its VEV determines the string coupling: $g_s = e^{\langle \Phi \rangle}$.

Total massless states (in $D = 26$): $24 \times 24 = 576$ physical degrees of freedom, decomposing as $299 + 276 + 1 = 576$ ($G + B + \Phi$).

Level $N = \bar{N} = 2$ (first massive level). $m^2 = 4/\alpha'$. The states are constructed from $\alpha_{-2}^\mu \bar{\alpha}_{-2}^\nu |0; p\rangle$, $\alpha_{-1}^\mu \alpha_{-1}^\rho \bar{\alpha}_{-2}^\sigma |0; p\rangle$, etc. The degeneracy grows rapidly: at level N , the number of states is $\sim p(N)^2$ (the square of the partition function of N), which grows as $e^{4\pi\sqrt{N}}$ for large N .

The superstring spectrum and the GSO projection

The superstring adds worldsheet fermions $\psi^\mu(z)$ with two possible boundary conditions:

Ramond (R): $\psi^\mu(\sigma + 2\pi) = +\psi^\mu(\sigma)$ (periodic). The R sector contains spacetime fermions.

Neveu–Schwarz (NS): $\psi^\mu(\sigma + 2\pi) = -\psi^\mu(\sigma)$ (antiperiodic). The NS sector contains spacetime bosons.

The closed superstring has four sectors: NS-NS, NS-R, R-NS, R-R (independently choosing boundary conditions for left- and right-movers).

The GSO projection. The Gliozzi–Scherk–Olive (GSO) projection keeps only states with a definite worldsheet fermion number parity $(-1)^F$: states with $(-1)^F = +1$ are kept; states with $(-1)^F = -1$ are projected out.

The GSO projection:

- (i) Removes the tachyon ($N = 0$ in the NS sector has $(-1)^F = -1$).
- (ii) Ensures spacetime supersymmetry (the remaining spectrum forms complete SUSY multiplets).
- (iii) Ensures modular invariance of the one-loop partition function (the different spin structures on the torus contribute with relative signs determined by the GSO projection).

The massless superstring spectrum. After the GSO projection:

NS-NS sector: $G_{\mu\nu}$ (graviton), $B_{\mu\nu}$ (Kalb–Ramond), Φ (dilaton) — the same bosonic fields as in the bosonic string, now with $8 \times 8 = 64$ physical degrees of freedom in $D = 10$.

R-R sector: p -form potentials C_p . Type IIA: C_1, C_3 (odd forms). Type IIB: C_0, C_2, C_4 (even forms). These are the fields that couple to D-branes (section 45.12).

NS-R and R-NS sectors: the gravitinos ψ_μ^α (the superpartners of the graviton) and the dilatinos (the superpartners of the dilaton).

Why the graviton is universal

The graviton appears in every closed string theory for a simple reason: the operator $\alpha_{-1}^\mu \bar{\alpha}_{-1}^\nu$ always exists (it is the first excited level of the universal oscillator algebra (45.25)), and it always gives a massless spin-2 state (by the mass-shell condition (45.27) with $N = \bar{N} = 1$ and the normal-ordering constant $a = 1$ or $1/2$).

No other known theory has this property: a massless spin-2 state as an automatic consequence of the algebraic structure. In QFT: a massless spin-2 field must be postulated (the linearised Einstein action is an input). In string theory: it is an output of the oscillator algebra.

By the Weinberg–Witten theorem and the analysis of Boulware and Deser: a consistent interacting massless spin-2 field must couple universally to the energy-momentum tensor and must satisfy the Einstein equation at low energies. The string graviton satisfies both conditions: its vertex operator V_{graviton} couples to $T^{\mu\nu}$ (by the worldsheet Ward identity), and its tree-level scattering amplitude reproduces the Einstein–Hilbert action (section 45.7).

The graviton is not put in by hand. It emerges from the oscillator algebra of the closed string, and the Einstein equation emerges from its scattering amplitudes. This is the strongest evidence that string theory contains gravity.

45.7 Scattering amplitudes: from worldsheet to spacetime

String perturbation theory computes scattering amplitudes by summing over worldsheet topologies. Each topology contributes a finite-dimensional integral over the moduli space of Riemann surfaces, weighted by the worldsheet path integral. The result is UV-finite at every order — the central computational miracle of string theory.

The genus expansion

The n -point scattering amplitude of closed strings:

$$\mathcal{A}_n = \sum_{g=0}^{\infty} g_s^{2g-2+n} \int_{\mathcal{M}_{g,n}} d^d_g \tau \left\langle \prod_{i=1}^n V_i(k_i, z_i) \right\rangle_{\Sigma_g}, \quad (45.28)$$

where g is the genus (number of handles), $g_s = e^{\langle \Phi \rangle}$ is the string coupling, $\mathcal{M}_{g,n}$ is the moduli space of Riemann surfaces of genus g with n punctures, τ are the moduli (the complex structure parameters), V_i are the vertex operators for the external states, and the expectation value is computed in the worldsheet CFT on Σ_g .

The moduli space has complex dimension $d_g = 3g - 3 + n$ (for $g \geq 2$), with special cases $d_0 = n - 3$ (sphere) and $d_1 = n$ (torus).

The genus expansion is the string-theory analogue of the Feynman-diagram loop expansion:

Order	Worldsheet	Weight	QFT analogue
Tree ($g = 0$)	Sphere	g_s^{n-2}	Tree-level Feynman diagrams
One loop ($g = 1$)	Torus	g_s^n	One-loop diagrams
g loops	Genus- g surface	g_s^{2g+n-2}	g -loop diagrams

The crucial difference: in QFT, each Feynman diagram is a separate integral, and the number of diagrams grows factorially with the loop order. In string theory, all diagrams at a given order are packaged into a *single* integral over the moduli space $\mathcal{M}_{g,n}$. The moduli-space integral replaces the sum over Feynman topologies: the string “knows” about all possible interaction histories simultaneously.

Tree-level: the Virasoro–Shapiro amplitude

The tree-level four-point amplitude of closed strings (genus $g = 0$, worldsheet = sphere) is the Virasoro–Shapiro amplitude:

$$\mathcal{A}_4^{(0)}(s, t, u) = -\frac{8\pi g_s^2}{\alpha'^3} \frac{\Gamma(-\alpha' s/4) \Gamma(-\alpha' t/4) \Gamma(-\alpha' u/4)}{\Gamma(1 + \alpha' s/4) \Gamma(1 + \alpha' t/4) \Gamma(1 + \alpha' u/4)}, \quad (45.29)$$

where s, t, u are the Mandelstam variables ($s + t + u = 0$ for massless external states) and $\alpha' = 1/(2\pi\sigma)$ is the Regge slope.

Properties of the Virasoro–Shapiro amplitude:

Crossing symmetry. $\mathcal{A}_4$ is symmetric under $s \leftrightarrow t \leftrightarrow u$: the amplitude is manifestly crossing-symmetric (invariant under exchange of any pair of external legs). In QFT: crossing symmetry is a consequence of summing over s -, t -, and u -channel Feynman diagrams. In string theory: it is automatic from the single moduli-space integral.

Regge poles. The gamma function $\Gamma(-\alpha' s/4)$ has poles at $\alpha' s/4 = 0, 1, 2, \dots$, i.e., $s = 4n/\alpha'$: these are the masses of the closed string excitations at level n . The residue at each pole is a

polynomial in t , corresponding to the exchange of a particle of spin $J \leq 2n$ on the n -th Regge trajectory.

Low-energy limit. At $s, t, u \ll 1/\alpha'$ (energies well below the string scale): $\Gamma(-x) \approx -1/x + \gamma_E + \dots$ for small x , and the amplitude reduces to:

$$\mathcal{A}_4^{(0)} \xrightarrow{\alpha' \rightarrow 0} \frac{8\pi G}{stu} K(s, t, u), \quad (45.30)$$

where $G = g_s^2 \alpha'^4 / (8\pi)$ (Newton's constant in string theory) and K is the kinematic factor of the graviton scattering amplitude. This is the tree-level result of general relativity: the string amplitude reproduces Einstein gravity at low energies.

High-energy behaviour: exponential softening

At high energies ($s \gg 1/\alpha'$, fixed angle θ): the Virasoro–Shapiro amplitude falls off exponentially:

$$\mathcal{A}_4^{(0)} \sim \exp\left(-\frac{\alpha'}{2}(s \ln s + t \ln t + u \ln u)\right) \sim e^{-\alpha' s f(\theta)}, \quad (45.31)$$

where $f(\theta) > 0$ is an angle-dependent function. This is the *Gross–Mende* behaviour (1988): string amplitudes fall off exponentially at high energies and fixed angle.

The contrast with point-particle QFT is dramatic:

	Point particle (QFT)	String
High-energy, fixed angle	Power-law: $\mathcal{A} \sim s^J$ (spin- J exchange)	Exponential falloff: $\mathcal{A} \sim e^{-\alpha' s f(\theta)}$
UV behaviour	Divergent loop integrals	Finite moduli-space integrals
Hard scattering	Point-like cross section	Exponentially suppressed

The exponential softening means strings cannot be probed at distances shorter than $\ell_s = \sqrt{\alpha'}$: the cross section for resolving sub-string structure falls off as $\sigma \sim e^{-E^2 \alpha'}$ with the centre-of-mass energy. The string length ℓ_s is a fundamental resolution limit, not because of a lattice or a cutoff, but because the extended nature of the string prevents hard point-like scattering.

The physical origin. At high energies, the dominant contribution to the scattering amplitude comes from a saddle point of the worldsheet path integral (the Gross–Mende saddle): a classical string configuration that spreads over a region of size $\sim \sqrt{\alpha' s}$. As the energy increases, the string spreads further, diluting the interaction over a larger area and suppressing the amplitude. The exponential falloff is the spreading of the string's spatial extent with energy.

45.8 UV finiteness: modular invariance

The central computational result of string theory: every order in the genus expansion (45.28) is UV-finite. No renormalisation is needed. No counterterms are introduced. No new parameters appear at high energies. We explain the mechanism.

The one-loop amplitude: the torus

The one-loop amplitude ($g = 1$) is an integral over the moduli space $\mathcal{M}_1$ of the torus. A torus is defined by a complex modular parameter $\tau = \tau_1 + i\tau_2$ (with $\tau_2 > 0$), and the moduli space is:

$$\mathcal{M}_1 = \mathbb{H}/SL(2, \mathbb{Z}), \quad (45.32)$$

where $\mathbb{H}$ is the upper half-plane and $SL(2, \mathbb{Z})$ is the modular group generated by $T : \tau \rightarrow \tau + 1$ and $S : \tau \rightarrow -1/\tau$.

The fundamental domain $\mathcal{F}$ of the modular group is:

$$\mathcal{F} = \{\tau \in \mathbb{H} : |\tau_1| \leq \frac{1}{2}, |\tau| \geq 1\}. \quad (45.33)$$

This is a finite region in the upper half-plane, bounded from below by $\tau_2 \geq \sqrt{3}/2$ (at $\tau_1 = 0$, $|\tau| = 1$).

The one-loop vacuum amplitude (the cosmological constant at one loop):

$$\mathcal{A}_1^{(0)} = \int_{\mathcal{F}} \frac{d^2\tau}{\tau_2^2} \frac{1}{\tau_2^{D/2}} Z(\tau, \bar{\tau}), \quad (45.34)$$

where $Z(\tau, \bar{\tau})$ is the worldsheet partition function (the trace over the string Hilbert space with weight $q^{L_0} \bar{q}^{\bar{L}_0}$, $q = e^{2\pi i \tau}$).

The UV/IR connection

In a point-particle QFT: the one-loop diagram is an integral over the loop momentum k , and the UV divergence comes from the $|k| \rightarrow \infty$ region. In string theory: the one-loop amplitude is an integral over τ , and the potential UV divergence would come from the $\tau_2 \rightarrow 0$ region (a thin, degenerate torus, which looks like a point-particle loop with high loop momentum).

But $\tau_2 \rightarrow 0$ is *not in the fundamental domain $\mathcal{F}$* : the modular transformation $S : \tau \rightarrow -1/\tau$ maps $\tau_2 \rightarrow 0$ to $\tau_2 \rightarrow \infty$ (the large-torus limit, which is the IR, not the UV). The UV region of the moduli space is *identified with the IR region* by the modular group.

This is the mechanism:

$$\text{UV (small } \tau_2) \xrightarrow{S: \tau \rightarrow -1/\tau} \text{IR (large } \tau_2). \quad (45.35)$$

The modular invariance of the string worldsheet — the requirement that the physics is independent of how the torus is parametrised — removes the UV region from the integration domain entirely. There is no UV divergence because there is no UV region to integrate over.

The point-particle comparison. In QFT, the one-loop integral in the Schwinger proper-time representation: $\Gamma^{(1)} = -\frac{1}{2} \int_0^\infty \frac{ds}{s} \text{Tr} e^{-sD}$, where s is the proper time (the circumference of the loop). The UV divergence comes from $s \rightarrow 0$ (small loops, high momenta).

In string theory, the proper time s is replaced by the modular parameter $\tau_2 \propto s$, and the integral runs over $\mathcal{F}$ instead of $(0, \infty)$. The key difference: the lower bound of τ_2 in $\mathcal{F}$ is $\tau_2 \geq \sqrt{3}/2$, not $\tau_2 \geq 0$. The modular group has cut out the $\tau_2 \rightarrow 0$ corner.

	Point particle	String
One-loop diagram	Circle (S^1)	Torus (T^2)
Integration variable	Schwinger time $s \in (0, \infty)$	Modular parameter $\tau \in \mathcal{F}$
UV region	$s \rightarrow 0$ (included $\rightarrow$ divergence)	$\tau_2 \rightarrow 0$ (excluded by S -duality)
UV divergence	$\int_0 ds/s^{1+D/2} = \infty$	$\int_{\mathcal{F}} d^2\tau/\tau_2^{2+D/2}$: finite

Higher loops and all-order finiteness

The argument extends to all genera. At genus g : the moduli space $\mathcal{M}_g$ is a finite-dimensional orbifold (the Teichmüller space modulo the mapping class group). The potential UV divergences

correspond to degenerate limits of the Riemann surface (pinched handles, nodes), which are on the boundary $\partial\mathcal{M}_g$. These boundary contributions either cancel between bosons and fermions (in the superstring: the GSO projection removes the tachyon, and spacetime supersymmetry ensures cancellation) or are identified with IR effects by modular invariance.

The result: string perturbation theory is UV-finite at every order in g_s . The genus expansion is an asymptotic series (like the QFT loop expansion), and the sum over genera may diverge, but each term is finite. No counterterms, no new parameters, no renormalisation.

The lattice-first perspective on UV completion

The lattice-first viewpoint (section ??) and string theory's UV completion offer two different answers to the same question: what happens at short distances?

The lattice-first answer: the short-distance structure is the lattice. The lattice spacing a is the engine's resolution floor. There is no physics below a . UV completion is not needed because there is no UV to complete: the engine is discrete.

The string answer: the short-distance structure is the string. The string length ℓ_s is the resolution floor. Sub-string distances are not probed because the cross section falls off exponentially. UV completion is achieved by modular invariance, which identifies UV and IR.

The two answers are not contradictory; they are complementary.

On a lattice with spacing a : the gauge fields live on links, the fermions on sites, and the path integral is a finite-dimensional integral. No UV divergences arise because the momentum is bounded by π/a . The lattice provides UV finiteness by brute force: there are no modes above the cutoff.

In string theory: the fields live on a continuous spacetime, but the scattering amplitudes are softened by the extended nature of the string. The modular invariance of the worldsheet provides UV finiteness by geometry: the UV region is identified with the IR.

If the lattice is the Fock-layer engine (as this book proposes), then the string's UV finiteness is a property of the rendering: the Born-layer amplitudes are UV-finite because the engine has no modes above π/a . The string's modular invariance is a rendered symmetry of the low-energy effective theory, not a property of the engine. The lattice provides the UV finiteness; the string explains why the rendered amplitudes are smooth.

This is speculative: whether string theory is the correct effective description of a lattice engine at $a \sim \ell_P$ is an open question. But the lattice-first framework is compatible with string-theoretic UV finiteness: both produce finite answers, by different mechanisms, at the same scale.

45.9 Einstein–Hilbert from worldsheet consistency

The sigma model (45.3) is a 2D QFT with coupling constants $G_{\mu\nu}(X)$, $B_{\mu\nu}(X)$, $\Phi(X)$. Weyl invariance at the quantum level requires the vanishing of the beta functions for these couplings:

$$\beta_{\mu\nu}^G = \alpha' \left(R_{\mu\nu} + 2\nabla_\mu \nabla_\nu \Phi - \frac{1}{4} H_{\mu\rho\sigma} H_\nu{}^{\rho\sigma} \right) + O(\alpha'^2) = 0, \quad (45.36)$$

$$\beta_{\mu\nu}^B = \alpha' \left(-\frac{1}{2} \nabla^\rho H_{\rho\mu\nu} + \nabla^\rho \Phi H_{\rho\mu\nu} \right) + O(\alpha'^2) = 0, \quad (45.37)$$

$$\beta^\Phi = \alpha' \left(\frac{D-26}{6\alpha'} - \frac{1}{2} \nabla^2 \Phi + \nabla_\mu \Phi \nabla^\mu \Phi - \frac{1}{24} H_{\mu\nu\rho} H^{\mu\nu\rho} \right) + O(\alpha'^2) = 0, \quad (45.38)$$

where $H_{\mu\nu\rho} = 3\partial_{[\mu} B_{\nu\rho]}$ is the field strength of the Kalb–Ramond field.

The equation $\beta_{\mu\nu}^G = 0$ at leading order in α' is $R_{\mu\nu} + 2\nabla_\mu \nabla_\nu \Phi = 0$: the Einstein equation (with dilaton source), derived not from the Einstein–Hilbert action but from the requirement that

the worldsheet CFT be consistent. The $\beta^B = 0$ equation gives the equation of motion for $B_{\mu\nu}$; $\beta^\Phi = 0$ gives the dilaton equation and fixes $D = 26$ (or $D = 10$ for the superstring).

General relativity is not postulated; it is derived as the consistency condition of the string worldsheet.

45.10 The Kalb–Ramond field

The antisymmetric tensor $B_{\mu\nu}(X)$ (Kalb–Ramond field) is the string-theoretic generalisation of the electromagnetic potential A_μ . The gauge potential A_μ couples to a point particle’s worldline ($\int A_\mu dx^\mu$); the Kalb–Ramond field couples to the string’s worldsheet ($\int B_{\mu\nu} dX^\mu \wedge dX^\nu$).

The field strength $H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} + \text{cyclic}$ is the analogue of $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The gauge invariance $B_{\mu\nu} \rightarrow B_{\mu\nu} + \partial_\mu \Lambda_\nu - \partial_\nu \Lambda_\mu$ (with gauge parameter Λ_μ) generalises $A_\mu \rightarrow A_\mu + \partial_\mu \Lambda$.

The B -field plays essential roles: it provides the torsion in the string effective action, it enters the anomaly cancellation mechanism (the Green–Schwarz mechanism modifies H to include Chern–Simons terms), and it participates in T-duality (under which $B_{\mu\nu}$ components mix with the metric).

45.11 Perturbative definitions and the search for M-theory

String theory, as developed in the previous subsections, is defined perturbatively: the genus expansion (45.72) computes scattering amplitudes as a power series in the string coupling g_s . Each term is well-defined and UV-finite, but the series itself is asymptotic (divergent). A non-perturbative definition — the analogue of the lattice for QFT — does not exist.

This is the deepest open problem in string theory: the theory is defined by its perturbative expansion, without a substrate from which the expansion is derived.

The five perturbative string theories

In ten spacetime dimensions, there are exactly five consistent superstring theories, each defined by its worldsheet CFT and its spectrum of massless fields:

Theory	Strings	SUSY	Gauge group	Massless bosonic content
Type IIA	Closed	$\mathcal{N} = (1, 1)$	None	$g_{\mu\nu}, B_{\mu\nu}, \Phi, C_1, C_3$
Type IIB	Closed	$\mathcal{N} = (2, 0)$	None	$g_{\mu\nu}, B_{\mu\nu}, \Phi, C_0, C_2, C_4$
Type I	Open Closed	+ $\mathcal{N} = 1$	$SO(32)$	$g_{\mu\nu}, B_{\mu\nu}, \Phi, A_\mu^a$
Het. $SO(32)$	Closed	$\mathcal{N} = 1$	$SO(32)$	$g_{\mu\nu}, B_{\mu\nu}, \Phi, A_\mu^a$
Het. $E_8 \times E_8$	Closed	$\mathcal{N} = 1$	$E_8 \times E_8$	$g_{\mu\nu}, B_{\mu\nu}, \Phi, A_\mu^a$

Each theory has its own perturbative expansion (its own genus expansion with its own worldsheet CFT), and each is internally consistent: anomaly-free, UV-finite at every genus, and unitary. In the 1980s, these five theories were thought to be distinct, competing candidates for a “theory of everything.”

The web of dualities

In the 1990s, a network of dualities was discovered connecting the five theories:

T-duality (section ??): compactify one dimension on a circle of radius R . T-duality maps $R \rightarrow \alpha'/R$, exchanging momentum modes (n/R) and winding modes (wR/α'). It maps Type

$\text{IIA} \leftrightarrow \text{Type IIB}$ and $\text{Het. } SO(32) \leftrightarrow \text{Het. } E_8 \times E_8$.

S-duality: maps strong coupling to weak coupling, $g_s \rightarrow 1/g_s$. It maps $\text{Type I} \leftrightarrow \text{Het. } SO(32)$ (at strong coupling, the Type I open strings become the heterotic closed strings) and maps Type IIB to itself ($g_s \rightarrow 1/g_s$: the theory is self-dual).

Type IIA at strong coupling: as $g_s \rightarrow \infty$, Type IIA develops an eleventh dimension of radius $R_{11} = g_s \ell_s$. The ten-dimensional theory “decompactifies” into an eleven-dimensional theory: M-theory.

Het. $E_8 \times E_8$ at strong coupling: decompactifies to M-theory on an interval $S^1/\mathbb{Z}_2$ (Hořava–Witten theory, 1996).

The five string theories are not five different theories. They are five perturbative expansions of a single underlying theory, valid in different regions of the moduli space (the space of couplings and compactification parameters). Each perturbative expansion is a Taylor series around a specific point in moduli space ($g_s \rightarrow 0$ for a specific theory), and the dualities connect the different expansion points.

M-theory: the conjectured unification

Witten (1995) proposed that the five string theories, together with eleven-dimensional supergravity, are limiting cases of a single theory in eleven dimensions: *M-theory*.

What is known about M-theory:

The low-energy limit: 11D supergravity, the unique maximally supersymmetric ($\mathcal{N} = 1$ in 11D, 32 supercharges) gravitational theory in eleven dimensions. The bosonic fields are the metric g_{MN} ($M, N = 0, \dots, 10$) and a 3-form potential C_{MNP} .

The extended objects: M2-branes (2+1 dimensional) and M5-branes (5+1 dimensional), which couple electrically and magnetically to C_3 . No fundamental string: the string of Type IIA is an M2-brane wrapped on the compactification circle S^1 ($R_{11} = g_s \ell_s$), and its tension $T_{\text{string}} = R_{11}/(2\pi\alpha') = 1/(2\pi\ell_s^2)$ is the M2-brane tension times the circumference.

The compactifications: M-theory on S^1 gives Type IIA. M-theory on $S^1/\mathbb{Z}_2$ gives Het. $E_8 \times E_8$. The other three string theories are obtained by further dualities and compactifications.

The BFSS matrix model: Banks, Fischler, Shenker, and Susskind (1997) proposed that M-theory in the infinite momentum frame (the discrete light-cone quantisation, DLCQ) is described by the quantum mechanics of $N \times N$ Hermitian matrices in the limit $N \rightarrow \infty$:

$$H_{\text{BFSS}} = \frac{1}{2} \text{Tr}(\dot{X}_i^2 - \frac{1}{4} [X_i, X_j]^2) + \text{fermionic terms}, \quad (45.39)$$

where X_i ($i = 1, \dots, 9$) are $N \times N$ matrices representing the transverse coordinates of N D0-branes. The off-diagonal entries describe the interactions between D0-branes; the diagonal entries describe their positions.

What is NOT known about M-theory

The fundamental degrees of freedom. In QFT: the fundamental degrees of freedom are fields $\phi(x)$, defined at each spacetime point. In perturbative string theory: the fundamental degree of freedom is the worldsheet embedding $X^\mu(\sigma, \tau)$. In M-theory: the fundamental degrees of freedom are unknown.

The M2-brane and M5-brane are extended objects, but they are not “fundamental” in the way that the string is fundamental in perturbative string theory: there is no worldvolume CFT

for the M5-brane (the $(2, 0)$ superconformal field theory in 6D is not a free theory and has no Lagrangian description), and the M2-brane worldvolume theory (the BLG/ABJM model) was constructed only in 2008, thirteen years after M-theory was proposed.

The action. There is no action for M-theory. The 11D supergravity action is the low-energy limit, but the full action (including all the higher-derivative corrections and the non-perturbative effects) is unknown. String field theory (section 45.16) provides an action for perturbative string theory, but it does not extend to M-theory (there is no “M-theory field theory”).

The Hilbert space. The Hilbert space of M-theory is unknown. The BFSS matrix model provides a Hilbert space (the space of $N \times N$ matrix quantum mechanics states, in the limit $N \rightarrow \infty$), but this is the Hilbert space in the DLCQ frame, not a covariant Hilbert space. Whether the BFSS model captures all of M-theory (and not just a specific kinematic limit) is an open question.

A non-perturbative definition. The most direct statement: there is no non-perturbative definition of M-theory from which the five perturbative string theories, the dualities, and 11D supergravity can be derived. What exists is:

- (i) Five perturbative expansions (the five string theories), each valid at weak coupling.
- (ii) A network of dualities connecting them, providing evidence that a single theory underlies all five.
- (iii) The low-energy limit (11D supergravity), which captures the classical dynamics but misses quantum effects.
- (iv) Partial non-perturbative definitions in specific backgrounds: the BFSS matrix model (in the DLCQ frame), the IKKT matrix model (a Euclidean 0-dimensional model), and AdS/CFT (which defines string/M-theory non-perturbatively in asymptotically AdS backgrounds).

None of these is a complete, background-independent, non-perturbative definition of M-theory in the same sense that the lattice partition function $Z = \int \prod dU \det D e^{-S_G}$ is a complete, background-independent, non-perturbative definition of QCD.

Why the non-perturbative definition has not been found

The difficulty is not merely technical. There are structural obstacles:

1. What to put on a lattice? In QFT: the fields $\phi(x)$ live on spacetime points, and the lattice is a discretisation of spacetime. The lattice defines the theory: the continuum theory is the rendering.

In string theory: the fundamental object is one-dimensional (the string) or higher-dimensional (branes). Discretising the target spacetime is straightforward, but it is not clear what the lattice degrees of freedom are. Lattice strings? Lattice branes? Lattice matrices? The answer depends on knowing the fundamental degrees of freedom of M-theory, which is precisely what is unknown.

2. The string field theory is non-local. The star product in Witten’s SFT involves all Fourier modes of the string simultaneously. The string field $\Phi[X(\sigma)]$ is a functional of the entire string configuration, not a field at a spacetime point. Lattice regularisation works for local field theories (the action couples only neighbouring sites), but the string field action is non-local in target space. A lattice regularisation of SFT would require discretising the space of string configurations — an infinite-dimensional function space.

3. Eleven dimensions and the M5-brane. M-theory is eleven-dimensional, and its extended objects include the M5-brane. The worldvolume theory of the M5-brane is a 6D $(2, 0)$

superconformal field theory — a theory with no known Lagrangian and no free-field limit. Any non-perturbative definition of M-theory must contain this theory as a subsector. But a theory with no Lagrangian cannot be discretised by standard lattice methods (which require an action to define the Boltzmann weight e^{-S}).

4. Supersymmetry on the lattice. M-theory has maximal supersymmetry (32 supercharges). Supersymmetry is notoriously difficult to preserve on a lattice: the SUSY algebra $\{Q, \bar{Q}\} = H + P$ relates the Hamiltonian to the supercharge, and the lattice discretisation of P (spatial momentum) breaks part of the SUSY algebra. Preserving some supercharges on the lattice is possible (using twisted formulations or orbifold constructions), but preserving all 32 is an open problem.

5. The dualities are non-perturbative. The dualities connecting the five string theories (S-duality, U-duality) map $g_s \rightarrow 1/g_s$: they exchange strong and weak coupling. A non-perturbative definition must contain all five perturbative limits simultaneously and reproduce the dualities as exact symmetries. This is a strong constraint: the definition must be self-dual under transformations that exchange its different perturbative expansions. No known mathematical framework provides this naturally.

The Fock-Born reading

The situation has a precise Fock-Born diagnosis:

The five string theories are Born-layer descriptions. Each perturbative string theory is an effective description valid at weak coupling ($g_s \ll 1$) in a specific corner of the moduli space. It is a rendering: it describes the scattering amplitudes, the spectrum, and the low-energy dynamics in a specific regime.

M-theory is the conjectured Fock-layer engine. The dualities provide evidence that a single theory underlies all five renderings. This theory — M-theory — would be the engine from which all five perturbative expansions are derived as Born-layer outputs in different limits.

The engine has not been found. The Fock-layer Hamiltonian of M-theory is unknown. What exists are five perturbative renderings (the string theories), a network of consistency conditions (the dualities), and a low-energy limit (11D supergravity). The situation is analogous to knowing the hadron spectrum and the scattering amplitudes without knowing QCD: the Born-layer data is abundant, but the Fock-layer engine is missing.

The BFSS matrix model is the closest candidate: it provides a Hamiltonian H_{BFSS} (45.39) acting on a Hilbert space of matrix quantum mechanics states. If BFSS is correct: M-theory IS a quantum mechanics of matrices, the fundamental degrees of freedom are $N \times N$ Hermitian matrices X_i ($i = 1, \dots, 9$), and the engine is a $(0 + 1)$ -dimensional theory (quantum mechanics, not field theory). Spacetime, strings, branes, and gravity would all be Born-layer renderings of matrix quantum mechanics — the most radical version of emergent spacetime.

Whether BFSS captures all of M-theory, or only a specific kinematic limit, remains the central open question. The absence of a complete non-perturbative definition is, in the Fock-Born framework, the absence of the engine. The renderings are known; the engine is not.

45.12 D-branes

D-branes are the most important non-perturbative objects in string theory. They are dynamical p -dimensional hypersurfaces on which open string endpoints are constrained to lie. They were introduced by Polchinski (1995) as the objects that carry Ramond–Ramond charge, and they are essential for the construction of realistic gauge theories, the formulation of AdS/CFT, and the understanding of black holes in string theory.

Boundary conditions and the definition

An open string has two endpoints ($\sigma = 0$ and $\sigma = \pi$). Each endpoint independently satisfies either Neumann or Dirichlet boundary conditions in each spacetime direction:

Neumann ($\partial_\sigma X^\mu|_{\sigma=0,\pi} = 0$): the endpoint is free to move in the μ -direction. No momentum flows off the string.

Dirichlet ($X^\mu|_{\sigma=0,\pi} = \text{const}$): the endpoint is fixed in the μ -direction. It is pinned to a specific value.

A Dp -brane is a $(p + 1)$ -dimensional hypersurface defined by imposing Neumann conditions in $p + 1$ directions (the worldvolume, including time) and Dirichlet conditions in the remaining $D - p - 1$ directions (the transverse space). The “D” stands for Dirichlet.

Object	p	Worldvol.	Description
D0-brane	0	$0 + 1$	Point particle; moves in time only
D1-brane	1	$1 + 1$	String-like object (“D-string”)
D2-brane	2	$2 + 1$	Membrane
D3-brane	3	$3 + 1$	Has a 4D worldvolume; central to AdS/CFT
Dp -brane	p	$p + 1$	General; p even in IIA, p odd in IIB

Not all values of p are allowed in a given string theory: Type IIA has D0, D2, D4, D6, D8; Type IIB has D1, D3, D5, D7, D9 (odd p only in IIB, even p only in IIA). This is dictated by the spectrum of Ramond–Ramond forms available in each theory.

Open string spectrum on a D-brane

The massless modes of an open string with both endpoints on the same Dp -brane form the worldvolume field content:

A gauge field $A_a(x)$ ($a = 0, 1, \dots, p$): a $U(1)$ gauge boson living on the $(p + 1)$ -dimensional worldvolume. This arises from the Neumann directions: the open string has a massless vector state from the NS sector, and its polarisation is along the worldvolume.

Scalar fields $\Phi^i(x)$ ($i = p + 1, \dots, D - 1$): one scalar per transverse direction. The scalar $\Phi^i(x)$ describes the displacement of the D-brane in the i -th transverse direction. Giving Φ^i a VEV moves the brane; making Φ^i fluctuate describes the brane’s shape oscillations. These are the Goldstone bosons of the broken translation symmetry (the brane breaks translation invariance in the transverse directions).

Fermions: superpartners of the gauge field and scalars (required by the worldvolume supersymmetry inherited from the bulk).

For a D3-brane in $D = 10$: the worldvolume is $3 + 1$ dimensional, with one $U(1)$ gauge field, $10 - 4 = 6$ real scalars, and their fermionic partners. This is the field content of $\mathcal{N} = 4$ $U(1)$ super Yang–Mills in $d = 4$.

Chan–Paton factors

Open string endpoints carry internal quantum numbers beyond their spacetime position and momentum. These are the *Chan–Paton factors* (Chan and Paton, 1969), originally introduced to attach flavour indices to the endpoints of open strings in dual resonance models, a decade before D-branes were understood.

The definition. Label the two endpoints of an open string by indices $i \in \{1, \dots, N\}$ (at $\sigma = 0$) and $j \in \{1, \dots, N\}$ (at $\sigma = \pi$). A general open string state is:

$$|\Psi\rangle = \sum_{i,j=1}^N \lambda_{ij} |\text{oscillator state}\rangle \otimes |i, j\rangle, \quad (45.40)$$

where λ_{ij} is an $N \times N$ matrix — the *Chan–Paton matrix* — and $|i, j\rangle$ specifies which labels sit on the two endpoints. The Chan–Paton indices are non-dynamical: they do not propagate along the string or contribute to the mass formula. They are labels, not fields.

In the D-brane interpretation (Polchinski 1995): the index i labels which D-brane the endpoint sits on. A string state $|i, j\rangle$ starts on brane i and ends on brane j . The Chan–Paton matrix λ_{ij} is the coefficient of this sector.

The structure of scattering amplitudes. The Chan–Paton factors factorise from the worldsheet computation. An n -point open string amplitude:

$$\mathcal{A}_n = \sum_{\text{orderings}} \text{Tr}(\lambda^{a_1} \lambda^{a_2} \dots \lambda^{a_n}) \times A_n^{\text{ws}}(k_1, \dots, k_n; \text{ordering}), \quad (45.41)$$

where λ^a are the Chan–Paton matrices of the external states, A_n^{ws} is the worldsheet amplitude (a function of the momenta and the cyclic ordering of the vertex operators on the boundary), and the sum is over inequivalent orderings.

The trace structure $\text{Tr}(\lambda^{a_1} \dots \lambda^{a_n})$ is the string-theoretic origin of the colour factor in gauge theory amplitudes. At low energies ($\alpha' \rightarrow 0$): the worldsheet amplitude reduces to the gauge theory amplitude, and the Chan–Paton trace reduces to the colour trace of the Feynman diagram. The colour-ordered decomposition of gauge theory amplitudes (section ??) is inherited from the Chan–Paton trace.

Gauge group determination. The Chan–Paton matrices λ^a are the generators of a Lie algebra. Which gauge group emerges depends on the type of string and the orientability of the worldsheet:

Oriented open strings (Type I', D-branes in Type II): both endpoints are distinguishable. The Chan–Paton matrices are general $N \times N$ matrices, forming the Lie algebra of $U(N)$. The gauge group is $U(N)$.

Unoriented open strings (Type I): the worldsheet orientation-reversal symmetry Ω identifies $|i, j\rangle \leftrightarrow \pm |j, i\rangle$. The Chan–Paton matrices must satisfy $\lambda^T = \pm \lambda$. Symmetric matrices give $SO(N)$; antisymmetric matrices give $USp(N)$.

In the Type I superstring: anomaly cancellation (the Green–Schwarz mechanism, 1984) uniquely selects $SO(32)$, fixing $N = 32$. This was the first indication that the gauge group of a string theory is not arbitrary but constrained by quantum consistency — the same structural miracle as anomaly cancellation in the Standard Model (section 41.4), now at the level of the string.

The $U(1)$ factor. For N D-branes in Type II: the gauge group is $U(N) = SU(N) \times U(1)$. The $U(1)$ factor (the overall trace $\lambda \propto \mathbf{1}_{N \times N}$) is the centre-of-mass $U(1)$ of the brane stack: it corresponds to the overall translational degree of freedom and typically decouples from the $SU(N)$ dynamics. For AdS/CFT purposes: the dual gauge theory is $SU(N)$ (the $U(1)$ decouples in the large- N limit).

The Fock-Born reading. Chan–Paton factors are the mechanism by which gauge symmetry emerges from the string-theoretic Fock layer. The fundamental degrees of freedom (open string oscillator modes) carry no gauge charge — they are neutral vibrations of a one-dimensional

object. The gauge structure is encoded entirely in the endpoint labels i, j : which brane the string connects. The N^2 labelling possibilities organise into the adjoint of $U(N)$, and the trace structure of amplitudes produces the colour factors of Yang–Mills theory.

In this sense, non-Abelian gauge symmetry is *twice emergent* in string theory: first, it emerges from the combinatorics of Chan–Paton labels (a labelling structure, not a dynamical symmetry of the worldsheet); second, it emerges at low energies when the massive string modes decouple and only the massless gauge bosons survive. The gauge group is a Born-layer rendering of the endpoint combinatorics of the Fock-layer engine.

Multiple D-branes and the emergence of gauge symmetry

The most remarkable property of D-branes is what happens when N of them are placed on top of each other.

Open strings now have N^2 sectors, labelled by which brane each endpoint sits on: the string can start on brane i and end on brane j , with $i, j = 1, \dots, N$. These are the *Chan–Paton factors*.

Diagonal strings ($i = j$): the string starts and ends on the same brane. There are N such strings, each giving a $U(1)$ gauge boson.

Off-diagonal strings ($i \neq j$): the string stretches between brane i and brane j . There are $N(N - 1)$ such strings (for $i \neq j$), and their massless modes are charged under both $U(1)_i$ and $U(1)_j$.

Together: the N^2 massless vector modes organise into the adjoint representation of $U(N)$. The gauge symmetry $U(N)$ is *emergent*: it arises from the combinatorics of open strings ending on N coincident branes, not from a postulate.

N	Gauge group	Worldvolume theory
1	$U(1)$	Abelian gauge theory + scalars
N coincident	$U(N)$	Non-Abelian gauge theory + adjoint scalars
N D3	$U(N)$	$\mathcal{N} = 4$ SYM in $d = 4$ (the AdS/CFT engine)

Separating the branes. If the N branes are separated in the transverse directions: $\Phi_{\text{VEV}}^i = \text{diag}(x_1^i, \dots, x_N^i)$, the gauge symmetry is broken. The off-diagonal strings stretching between separated branes acquire a mass proportional to the inter-brane distance: $m_{ij} = |x_i - x_j|/(2\pi\alpha')$. This is the string-theoretic realisation of the Higgs mechanism: the scalar VEV (brane separation) gives mass to the gauge bosons (the stretched strings), breaking $U(N) \rightarrow U(1)^N$.

The continuous moduli space of brane separations is the Coulomb branch of the gauge theory. The low-energy theory at a generic point on the Coulomb branch is $U(1)^N$ with massive W -bosons — exactly as in a spontaneously broken non-Abelian gauge theory.

D-brane charge and the WZ action

D-branes carry Ramond–Ramond (RR) charge. A Dp -brane is a source for the RR $(p + 1)$ -form potential C_{p+1} (just as an electron is a source for the electromagnetic 1-form potential A_μ). The charge:

$$\mu_p = T_p = \frac{1}{(2\pi)^p \alpha'^{(p+1)/2} g_s}, \quad (45.42)$$

where T_p is the brane tension (energy per unit worldvolume). The equality $\mu_p = T_p$ (charge equals tension) is required by BPS stability: the D-brane preserves half the supersymmetry, and the gravitational attraction is exactly cancelled by the RR repulsion. This is a no-force

condition: parallel BPS branes neither attract nor repel, which is why N coincident branes are a stable configuration.

The coupling to the RR field is described by the Wess–Zumino (WZ) action:

$$S_{\text{WZ}} = \mu_p \int_{\mathcal{W}} \sum_q C_q \wedge e^{B+2\pi\alpha' F}, \quad (45.43)$$

where $\mathcal{W}$ is the worldvolume, C_q are the RR potentials, B is the NS 2-form, F is the worldvolume gauge field strength, and the wedge product picks out the $(p+1)$ -form piece. The exponential $e^{B+2\pi\alpha' F}$ generates couplings to lower RR forms: a D p -brane with worldvolume flux $F \neq 0$ carries D $(p-2)$ -brane charge (“induced D-brane charge”).

The total D-brane action is the sum of the DBI and WZ pieces: $S = S_{\text{DBI}} + S_{\text{WZ}}$.

The DBI action

The kinetic part of the D-brane action is the Dirac–Born–Infeld (DBI) action (Dirac 1962, Born–Infeld 1934, adapted to D-branes by Leigh 1989):

$$S_{\text{DBI}} = -T_p \int d^{p+1}\xi \, e^{-\Phi} \sqrt{-\det(G_{ab} + B_{ab} + 2\pi\alpha' F_{ab})}, \quad (45.44)$$

where ξ^a ($a = 0, \dots, p$) are the worldvolume coordinates, $G_{ab} = \partial_a X^\mu \partial_b X^\nu g_{\mu\nu}$ is the pullback of the spacetime metric (the induced metric on the brane), B_{ab} is the pullback of the Kalb–Ramond 2-form, F_{ab} is the worldvolume $U(1)$ gauge field strength, and Φ is the dilaton.

The DBI action has two important limits:

Slow variation ($\alpha' F \ll 1$, small field strengths): expanding the square root gives $S_{\text{DBI}} \approx -T_p \int d^{p+1}\xi \sqrt{-G} [1 + \frac{1}{4}(2\pi\alpha')^2 F_{ab} F^{ab} + \dots]$. The leading term is the brane tension (cosmological constant on the worldvolume). The F^2 term is the Maxwell action: $\mathcal{L} \supset -\frac{1}{4g_{\text{YM}}^2} F^2$ with $g_{\text{YM}}^2 = (2\pi)^{p-2} \alpha'^{(p-3)/2} g_s$. For $p = 3$: $g_{\text{YM}}^2 = 2\pi g_s$, which is the YM coupling of the D3-brane gauge theory.

Static brane ($F_{ab} = 0$, $B_{ab} = 0$): the DBI action reduces to $S = -T_p \int d^{p+1}\xi \sqrt{-G}$, which is the Nambu–Goto action for a $(p+1)$ -dimensional extended object — the generalisation of the string action to higher-dimensional branes. The brane minimises its worldvolume (generalised area) just as a string minimises its worldsheet area.

D-brane intersections and gauge theory engineering

Different types of gauge theories can be *engineered* by arranging D-branes in specific configurations. The open strings stretching between different stacks of branes provide the matter content; the intersection geometry determines the gauge group and the representations:

Parallel D3-branes: N coincident D3-branes give $\mathcal{N} = 4$ $U(N)$ SYM — maximal supersymmetry in $d = 4$.

D3-branes at orbifold singularities: placing D3-branes at a $\mathbb{C}^2/\Gamma$ orbifold (where Γ is a discrete subgroup of $SU(2)$) breaks the supersymmetry to $\mathcal{N} = 2$ or $\mathcal{N} = 1$ and produces product gauge groups with bifundamental matter — the “quiver gauge theories” whose gauge group and matter content are encoded in the McKay quiver of Γ .

D3/D7 systems: D3-branes intersecting D7-branes add fundamental matter (strings stretching from D3 to D7 give quarks transforming in the fundamental of the D3 gauge group).

Hanany–Witten setups: D3-branes suspended between NS5-branes engineer $\mathcal{N} = 2$ theories in $d = 4$. The Coulomb branch is the moduli space of D3-brane positions between the NS5-branes; the Seiberg–Witten curve is the shape of the NS5-branes.

These constructions demonstrate that the entire zoo of four-dimensional gauge theories — different gauge groups, matter representations, amounts of supersymmetry — can be realised as different arrangements of D-branes in string theory. The gauge theory is not postulated; it *emerges* from the open string spectrum on the brane configuration.

T-duality and D-branes

T-duality (section ??) maps a string on a circle of radius R to a string on a circle of radius α'/R . Under T-duality in the X^9 direction:

Neumann $\leftrightarrow$ Dirichlet: a Dp -brane wrapped on the X^9 circle becomes a $D(p-1)$ -brane (the Neumann condition in X^9 becomes Dirichlet, and the brane no longer extends in that direction). Conversely, a Dp -brane transverse to X^9 becomes a $D(p+1)$ -brane (the Dirichlet condition becomes Neumann).

This is why IIA has even- p branes and IIB has odd- p branes: T-duality maps IIA to IIB, changing p by ± 1 .

T-duality also maps the gauge field on a brane to the brane position in the dual picture: $A_9 \leftrightarrow X^9/(2\pi\alpha')$. A Wilson line $\oint A_9 dx^9$ on the original brane becomes the transverse position of the dual brane. This is a concrete example of a geometric degree of freedom (position) emerging from a gauge degree of freedom (Wilson line) — or vice versa.

D-branes as black holes

At strong coupling ($g_s N \gg 1$), a stack of N D-branes curves the surrounding spacetime and looks like a black hole (or a “black brane”). The Strominger–Vafa calculation (1996): the Bekenstein–Hawking entropy of a five-dimensional extremal black hole (constructed from D1-branes, D5-branes, and momentum) matches *exactly* the microscopic state count from the D-brane worldvolume theory:

$$S_{\text{BH}} = \frac{A}{4G_5} = 2\pi\sqrt{N_1 N_5 n} = S_{\text{micro}}, \quad (45.45)$$

where N_1 , N_5 are the numbers of D1- and D5-branes and n is the momentum quantum number. This was the first statistical derivation of black hole entropy: the $e^{S_{\text{BH}}}$ microstates are counted from the open strings on the D-brane configuration, and the count matches the Bekenstein–Hawking area formula. This result provided strong evidence that string theory contains the correct microscopic degrees of freedom for quantum gravity.

The Fock-Born reading

D-branes are the clearest example of gauge symmetry as an emergent property of the Fock-layer engine:

The engine: N D-branes in flat spacetime with open strings. The fundamental degrees of freedom are the string modes, labelled by Chan–Paton indices (i, j) .

The rendering: at energies $E \ll 1/\sqrt{\alpha'}$ (below the string scale), the massive string modes decouple and the low-energy physics is described by a $U(N)$ gauge theory on the brane worldvolume. The gauge symmetry, the gauge bosons, the adjoint matter, and the gauge coupling constant all emerge from the open string spectrum.

The Higgs mechanism: separating branes is spontaneous symmetry breaking. The off-diagonal string mass $m_{ij} \propto |x_i - x_j|$ is the W -boson mass. The brane positions are the scalar VEVs on the Coulomb branch.

The black hole microstate count: the Strominger–Vafa formula (45.45) is a Fock-layer computation (counting open string states on the D-brane worldvolume) that matches a Born-layer observable (the Bekenstein–Hawking entropy of the macroscopic black hole). This is the gravitational analogue of the bridges from section 21.8: the engine’s microstate count matches the rendered thermodynamic entropy.

45.13 AdS/CFT: the holographic correspondence

The anti-de Sitter/conformal field theory correspondence (Maldacena 1997) is the most precise non-perturbative duality in theoretical physics. It states that a quantum field theory in d dimensions (without gravity) is exactly equivalent to a gravitational theory in $d + 1$ dimensions. The gravitational theory “lives in the bulk”; the field theory “lives on the boundary.” The extra dimension encodes the energy scale of the field theory: moving into the bulk corresponds to coarse-graining.

This section develops the correspondence in detail: the geometry of anti-de Sitter space, the D3-brane argument, the holographic dictionary, the computation of correlators, and the Ryu–Takayanagi formula for entanglement entropy.

Anti-de Sitter space

Anti-de Sitter space AdS_{d+1} is the maximally symmetric solution of the Einstein equation with a *negative* cosmological constant $\Lambda < 0$. In Poincaré coordinates:

$$ds^2 = \frac{L^2}{z^2}(-dt^2 + d\mathbf{x}^2 + dz^2), \quad (45.46)$$

where $L = \sqrt{-d(d-1)/(2\Lambda)}$ is the AdS radius and $z \in (0, \infty)$ is the radial coordinate. The boundary is at $z \rightarrow 0$ (“the UV”); the deep interior is at $z \rightarrow \infty$ (“the IR”).

Key features:

Conformal boundary. The metric (45.46) diverges as $z \rightarrow 0$: the boundary is at infinite proper distance but at finite conformal distance. Surfaces of constant z are copies of d -dimensional Minkowski space $\mathbb{R}^{d-1,1}$, scaled by L/z . The conformal boundary at $z = 0$ is where the dual CFT lives.

Isometry group. AdS_{d+1} has isometry group $SO(d, 2)$, which is the conformal group of d -dimensional Minkowski space. This is not a coincidence: it is the first entry in the holographic dictionary.

Negative curvature. The Ricci scalar is $R = -d(d+1)/L^2$: AdS is negatively curved, like a saddle. Geodesics oscillate rather than diverge (unlike de Sitter, where they fly apart). This makes AdS a “box”: massive particles in AdS are confined to the interior by the gravitational potential, and the boundary acts as a reflecting wall for normalisable modes.

The radial direction as energy scale. The metric factor L^2/z^2 means that a physical length ℓ at radial position z corresponds to a coordinate length $\ell_{\text{coord}} = \ell \cdot z/L$. Large z (deep interior) $\rightarrow$ large coordinate lengths $\rightarrow$ long wavelengths $\rightarrow$ low energies. Small z (near boundary) $\rightarrow$ short coordinate lengths $\rightarrow$ short wavelengths $\rightarrow$ high energies. The radial coordinate z is the *energy scale* of the dual field theory: the holographic RG.

The D3-brane argument

Consider N coincident D3-branes in type IIB string theory. The system has two descriptions, valid in different regimes:

Open string description (weak coupling: $g_s N \ll 1$): the open strings ending on the D3-branes reduce at low energies to $\mathcal{N} = 4$ super Yang–Mills (SYM) with gauge group $U(N)$ in $d = 4$. This is a conformal field theory: the beta function of $\mathcal{N} = 4$ SYM vanishes exactly at all orders (the coupling does not run).

Closed string description (strong coupling: $g_s N \gg 1$): the D3-branes source a geometry. The metric:

$$ds^2 = H^{-1/2}(r) \eta_{\mu\nu} dx^\mu dx^\nu + H^{1/2}(r) (dr^2 + r^2 d\Omega_5^2), \quad H(r) = 1 + \frac{L^4}{r^4}, \quad (45.47)$$

where $L^4 = 4\pi g_s N \alpha'^2$. In the near-horizon limit $r \ll L$ (equivalently $H \approx L^4/r^4$): the metric becomes $\text{AdS}_5 \times S^5$ with radius L :

$$ds^2 = \frac{r^2}{L^2} \eta_{\mu\nu} dx^\mu dx^\nu + \frac{L^2}{r^2} dr^2 + L^2 d\Omega_5^2. \quad (45.48)$$

Setting $z = L^2/r$ recovers the Poincaré form (45.46).

Maldacena’s insight: since the open and closed string descriptions describe the *same* system, the low-energy limits must be equivalent:

$$\boxed{\mathcal{N} = 4 \text{ SYM with } SU(N) \text{ in } d = 4 \longleftrightarrow \text{Type IIB on } \text{AdS}_5 \times S^5.} \quad (45.49)$$

The parameters

The duality maps the field-theory parameters to the gravitational parameters:

$$g_{\text{YM}}^2 = 4\pi g_s, \quad (45.50)$$

$$\lambda = g_{\text{YM}}^2 N = \frac{L^4}{\alpha'^2} \quad (\text{'t Hooft coupling}), \quad (45.51)$$

$$N = \frac{L^4}{\ell_s^4 4\pi g_s}. \quad (45.52)$$

Two limits are especially useful:

Large N , large λ ($N \rightarrow \infty, \lambda \rightarrow \infty$): the bulk is weakly curved ($L \gg \ell_s$: classical supergravity is valid) and weakly coupled ($g_s \ll 1$: no string loops). The dual field theory is at strong coupling. This is the power of the duality: strongly coupled gauge theory is computed by classical gravity.

Large N , small λ ($N \rightarrow \infty, \lambda \rightarrow 0$): the bulk is highly curved ($L \sim \ell_s$: stringy effects dominate). The dual field theory is at weak coupling. This regime is less useful (perturbation theory works on the field theory side), but it confirms the duality at weak coupling.

The limit $N \rightarrow \infty$ on the field theory side is the 't Hooft large- N expansion. In this limit: Feynman diagrams organise by topology (planar diagrams dominate, with corrections suppressed by $1/N^2$). On the gravity side: each topology corresponds to a genus of the string worldsheet. The $1/N^2$ expansion of the gauge theory *is* the genus expansion of string theory.

The holographic dictionary: GKPW

The precise formulation (Gubser, Klebanov, Polyakov 1998; Witten 1998):

$$\boxed{Z_{\text{CFT}}[J] = Z_{\text{gravity}}[\phi|_{z \rightarrow 0} = J].} \quad (45.53)$$

The generating functional of the CFT (left side) equals the gravitational partition function (right side), with the boundary value of each bulk field $\phi(z, x)$ set equal to the source $J(x)$ of the corresponding CFT operator.

Each bulk field corresponds to a CFT operator:

Bulk field	CFT operator	Relation
Metric $g_{\mu\nu}$	Stress tensor $T^{\mu\nu}$	Gravity is dual to energy-momentum
Scalar ϕ of mass m	Scalar $\mathcal{O}$ of dim Δ	$m^2 L^2 = \Delta(\Delta - d)$
Gauge field A_μ	Conserved current J^μ	Bulk gauge symmetry = boundary global symmetry
Gravitino ψ_μ	Supercurrent S^μ	Bulk SUSY = boundary SUSY

Holographic computation of correlators

In the large- N , large- λ limit: the gravity side is classical. The gravitational partition function is dominated by its saddle point (the classical solution): $Z_{\text{gravity}}[\phi|_{\partial} = J] \approx e^{-S_{\text{on-shell}}[J]}$.

The CFT generating functional: $W_{\text{CFT}}[J] = \ln Z_{\text{CFT}}[J] = -S_{\text{on-shell}}[J]$, and correlators follow by differentiation:

$$\langle \mathcal{O}(x_1) \cdots \mathcal{O}(x_n) \rangle = \frac{\delta^n (-S_{\text{on-shell}})}{\delta J(x_1) \cdots \delta J(x_n)} \Big|_{J=0}. \quad (45.54)$$

Worked example: the two-point function. Consider a massive scalar ϕ in AdS_{d+1} with action $S = \frac{1}{2} \int d^{d+1}x \sqrt{g} (g^{MN} \partial_M \phi \partial_N \phi + m^2 \phi^2)$. The equation of motion in the Poincaré metric (45.46):

$$z^{d+1} \partial_z (z^{1-d} \partial_z \phi) + z^2 \square_d \phi - m^2 L^2 \phi = 0, \quad (45.55)$$

where $\square_d$ is the d -dimensional d'Alembertian on the boundary. Near the boundary $z \rightarrow 0$, the solution behaves as:

$$\phi(z, x) \sim A(x) z^{d-\Delta} + B(x) z^\Delta, \quad \Delta = \frac{d}{2} + \sqrt{\frac{d^2}{4} + m^2 L^2}. \quad (45.56)$$

The leading mode $A(x)$ is the source (identified with $J(x)$); the subleading mode $B(x)$ is the response (the VEV $\langle \mathcal{O}(x) \rangle_J \propto B(x)$).

Evaluating the on-shell action and differentiating twice:

$$\langle \mathcal{O}(x) \mathcal{O}(0) \rangle = \frac{c_\Delta}{|x|^{2\Delta}}, \quad c_\Delta = \frac{\Gamma(\Delta)}{\pi^{d/2} \Gamma(\Delta - d/2)} \cdot \frac{(2\Delta - d)}{2}, \quad (45.57)$$

which is the two-point function of a conformal primary of dimension Δ — exactly the form required by conformal invariance. The holographic calculation produces the correct CFT answer from a classical gravity computation: solving a wave equation in one extra dimension.

Holographic thermodynamics

A black hole in the AdS bulk corresponds to a thermal state in the boundary CFT. The AdS-Schwarzschild metric:

$$ds^2 = \frac{L^2}{z^2} \left(-f(z) dt^2 + d\mathbf{x}^2 + \frac{dz^2}{f(z)} \right), \quad f(z) = 1 - \frac{z^4}{z_H^4}, \quad (45.58)$$

where z_H is the horizon location. The Hawking temperature: $T = 1/(\pi z_H)$ (in units with $L = 1$).

The free energy $F = -T \ln Z = T S_{\text{on-shell}}^E$ is computed from the Euclidean on-shell action. The result for $\mathcal{N} = 4$ SYM at strong coupling:

$$F = -\frac{\pi^2}{8} N^2 T^4 V_3. \quad (45.59)$$

At weak coupling (free field calculation): $F_{\text{free}} = -\frac{\pi^2}{6} N^2 T^4 V_3$. The ratio:

$$\frac{F_{\text{strong}}}{F_{\text{free}}} = \frac{3}{4}. \quad (45.60)$$

The strong-coupling free energy is 3/4 of the free-field value: the interactions reduce the effective number of degrees of freedom by 25%. This 3/4 factor was the first quantitative prediction of AdS/CFT (Gubser, Klebanov, Peet 1996), later confirmed by lattice calculations extrapolated to $\mathcal{N} = 4$.

The viscosity bound. The shear viscosity η of the strongly coupled plasma, computed holographically from the absorption cross section of gravitons by the AdS black hole:

$$\frac{\eta}{s} = \frac{1}{4\pi} \cdot \frac{\hbar}{k_B}, \quad (45.61)$$

where s is the entropy density. This is the Kovtun–Son–Starinets (KSS) bound, conjectured to be a universal lower bound for all fluids. The quark-gluon plasma at RHIC has $\eta/s \approx 1 - 2.5 \times \hbar/(4\pi k_B)$: the most perfect fluid ever observed, close to the holographic prediction.

45.14 The Ryu–Takayanagi formula

The Ryu–Takayanagi (RT) formula (2006) is the holographic computation of entanglement entropy. It is one of the most important results in theoretical physics of the 21st century: it connects geometry (minimal surfaces), information theory (entanglement entropy), and quantum gravity in a single equation.

The setup. Consider a CFT in its ground state, on a time slice. Divide the spatial region into a subregion A and its complement $\bar{A}$. The entanglement entropy of A :

$$S_A = -\text{Tr}_A(\hat{\rho}_A \ln \hat{\rho}_A), \quad \hat{\rho}_A = \text{Tr}_{\bar{A}} |\Psi\rangle \langle \Psi|, \quad (45.62)$$

where $\hat{\rho}_A$ is the reduced density matrix obtained by tracing over the complement $\bar{A}$. In a CFT, S_A diverges (UV divergence from the entanglement across the boundary ∂A) but its structure is universal.

The formula. In the holographic dual: the entanglement entropy of A is computed by finding the minimal-area surface γ_A in the bulk that is *homologous* to A (i.e., $\partial\gamma_A = \partial A$ and γ_A can be continuously deformed to A within the bulk):

$$S_A = \frac{\text{Area}(\gamma_A)}{4G_N}, \quad (45.63)$$

where G_N is Newton’s constant in the bulk. This is the Ryu–Takayanagi formula.

The formula is structurally identical to the Bekenstein–Hawking entropy $S_{\text{BH}} = A/(4G)$ — but applied to an arbitrary surface γ_A , not just a black hole horizon. The Bekenstein–Hawking entropy is the special case where A is the entire boundary (so γ_A is the black hole horizon).

Why it works. The RT formula can be understood from the structure of the holographic dictionary. The bulk geometry encodes the CFT’s correlators. The entanglement entropy S_A is

computed from the correlators of operators localised near ∂A (the twist operators of the replica trick). In the bulk, these twist operators correspond to cosmic branes that wrap the minimal surface γ_A . The on-shell action of the cosmic brane reduces to the area of γ_A in the classical limit — giving the RT formula.

A more direct argument: the Euclidean path integral for the n -th Rényi entropy $\text{Tr}(\rho_A^n)$ in the CFT maps to a gravitational path integral on a replica geometry (the bulk with a conical singularity at γ_A). The saddle point of this path integral localises on the minimal surface. Taking $n \rightarrow 1$ gives the von Neumann entropy $S_A = -\partial_n \ln \text{Tr}(\rho_A^n)|_{n=1} = \text{Area}(\gamma_A)/(4G_N)$.

Worked example: an interval in $\text{AdS}_3/\text{CFT}_2$. Consider a CFT_2 (a 2D conformal field theory) on a line, with subregion A an interval of length ℓ . The holographic dual is AdS_3 in Poincaré coordinates: $ds^2 = L^2(dz^2 + dx^2)/z^2$ (Euclidean, time suppressed).

The minimal “surface” (a geodesic in 3D) anchored at $\partial A = \{x = -\ell/2, x = +\ell/2\}$ on the boundary is a semicircle:

$$x^2 + z^2 = (\ell/2)^2. \quad (45.64)$$

Its length (regularised with a UV cutoff $z > \varepsilon$):

$$\text{Length}(\gamma_A) = 2L \int_{\varepsilon}^{\ell/2} \frac{dz}{z\sqrt{1 - (2z/\ell)^2}} = 2L \ln \frac{\ell}{\varepsilon}. \quad (45.65)$$

The RT formula (45.63) gives:

$$S_A = \frac{2L}{4G_N} \ln \frac{\ell}{\varepsilon} = \frac{c}{3} \ln \frac{\ell}{\varepsilon}, \quad (45.66)$$

where $c = 3L/(2G_N)$ is the central charge of the CFT_2 (the Brown–Henneaux formula). This matches the celebrated field-theory result of Calabrese and Cardy (2004) for the entanglement entropy of an interval in a 2D CFT: $S_A = (c/3) \ln(\ell/\varepsilon)$. The holographic calculation reproduces the exact CFT answer from a geodesic length.

Properties of the RT formula. The RT formula satisfies all the properties required of an entanglement entropy:

Strong subadditivity. $S_{A \cup B} + S_{A \cap B} \leq S_A + S_B$ for any regions A, B . This follows from the minimality of the RT surfaces: the minimal surface for $A \cup B$ can be deformed into surfaces that bound A and B separately, whose total area is at least $\text{Area}(\gamma_A) + \text{Area}(\gamma_B)$.

Area law. For a region A with smooth boundary ∂A in $d > 2$: $S_A = c_1 \text{Area}(\partial A)/\varepsilon^{d-2} + \dots$ (the leading divergence is proportional to the boundary area, not the volume). This follows from the RT formula because the minimal surface approaches the boundary, and its area near the boundary is dominated by the $1/z^{d-1}$ factor in the metric.

Mutual information. $I(A : B) = S_A + S_B - S_{A \cup B} \geq 0$, with a phase transition (the “entanglement wedge” transition) as the separation between A and B increases: the minimal surface for $A \cup B$ jumps from two disconnected surfaces (small separation) to one connected surface (large separation).

The quantum correction: the FLM and QES formulae. The RT formula is the leading (classical gravity) result. The quantum correction (Faulkner, Lewkowycz, Maldacena 2013):

$$S_A = \frac{\text{Area}(\gamma_A)}{4G_N} + S_{\text{bulk}}(\Sigma_A) + O(G_N), \quad (45.67)$$

where $S_{\text{bulk}}(\Sigma_A)$ is the von Neumann entropy of the bulk quantum fields in the “entanglement wedge” Σ_A (the bulk region bounded by A and γ_A). The *quantum extremal surface* (QES)

prescription (Engelhardt–Wall 2014) generalises this: minimise the *generalised entropy* $S_{\text{gen}} = \text{Area}(\gamma)/(4G_N) + S_{\text{bulk}}(\Sigma)$ over all surfaces γ homologous to A .

The QES formula is the tool that produces the Page curve for evaporating black holes (section 43.6): the “island” behind the horizon is the entanglement wedge of the radiation, and the QES jumps to include the island at the Page time, causing the entanglement entropy to decrease.

The Fock-Born reading of holography

AdS/CFT is the closest existing realisation of the Fock-Born architecture applied to gravity:

The CFT is the Fock-layer engine. The boundary CFT is a non-gravitational quantum field theory defined on flat spacetime. It has a Hilbert space, a Hamiltonian, a well-defined path integral, and no dynamical metric. It is a legitimate Fock-layer engine in the sense of this book: a quantum system evolving under $\dot{h}_t = -iH_{\text{CFT}}h_t$.

The AdS bulk is the Born-layer rendering. The bulk geometry — the metric g_{MN} , the matter fields ϕ , the black hole, the minimal surfaces — is the Born-layer output of the CFT engine. The geometry emerges from the CFT’s correlators via the holographic dictionary. The bulk metric is not quantised (in the classical gravity limit): it is a classical, macroscopic rendering of the CFT’s quantum state.

The holographic dictionary is the bridge formula. The GKPW relation (45.53) is the bridge from Fock-layer quantities ($Z_{\text{CFT}}[J]$, the correlators $\langle \mathcal{O} \rangle$) to Born-layer observables (the bulk geometry, the minimal-surface areas, the black hole thermodynamics).

The Ryu–Takayanagi formula is a Born-layer observable. S_A is the entanglement entropy of the CFT state restricted to region A : a Fock-layer quantity (a property of the reduced density matrix $\hat{\rho}_A$). The RT formula computes it from the area of a minimal surface in the bulk: a Born-layer geometric quantity. The formula $S_A = \text{Area}/(4G_N)$ is the bridge.

The one-way valve is respected. The CFT engine does not contain a dynamical metric: the CFT lives on flat space. The bulk metric g_{MN} is an output of the dictionary, not an input to the CFT Hamiltonian. AdS/CFT does not recycle the rendered geometry back into the engine. This is why it avoids the problem of time (the CFT has a perfectly good Hamiltonian H_{CFT} that generates unitary time evolution), the problem of observables (the CFT has standard Born-layer observables), and the non-renormalisability (the CFT is a well-defined quantum theory, with or without a continuum limit).

Fock-Born element	In AdS/CFT	Status
Fock-layer engine	CFT on the boundary	Non-gravitational, well-defined
Born-layer rendering	Bulk geometry	Emergent from CFT correlators
Bridge formula	GKPW dictionary (45.53)	Exact (in principle)
One-way valve	CFT $\rightarrow$ bulk, not bulk $\rightarrow$ CFT	Respected
Degeneracy $\rightarrow$ thermality	CFT thermal state = AdS black hole	Hawking temperature from CFT
Entanglement $\rightarrow$ geometry	RT formula (45.63)	$S_A = \text{Area}/(4G_N)$
Information paradox	Page curve from QES/islands	Unitarity preserved by CFT

AdS/CFT demonstrates that a non-gravitational engine can produce a gravitational rendering. The bulk spacetime — with its black holes, its gravitational waves, its Einstein equation — is not fundamental. It is the holographic encoding of the CFT’s quantum state. The geometry is

information; the information is quantum; the quantum theory is non-gravitational.

Whether this architecture extends beyond AdS (to cosmological spacetimes, to flat space, to the real universe with $\Lambda > 0$) is the central open question. The Fock-Born framework says it should: the engine should be non-gravitational, and the metric should be a Born-layer output. AdS/CFT is the proof of concept.

45.15 Tensor networks and the geometry of entanglement

The AdS/CFT correspondence (section 45.13) establishes the principle: a non-gravitational engine (the boundary CFT) renders a gravitational bulk (the AdS geometry). But the correspondence is a duality, not a construction: it tells us that the boundary theory *is equivalent to* the bulk, without showing how the geometry *emerges from* the boundary degrees of freedom.

Tensor networks provide the construction. They show, concretely and computably, how the geometry of the bulk spacetime is encoded in the entanglement structure of the boundary state — entirely within the Hamiltonian framework, with no path integral, no metric, and no graviton.

Tensor network states

A quantum state $|\Psi\rangle$ of N lattice sites, each with local dimension d , lives in a Hilbert space of dimension d^N . For $N \sim 100$ and $d = 2$: $d^N \sim 10^{30}$. Storing the full state vector is impossible.

A *tensor network state* approximates $|\Psi\rangle$ by decomposing the d^N -dimensional coefficient tensor into a network of smaller tensors connected by contracted indices:

$$\Psi_{s_1 s_2 \dots s_N} = \sum_{\{a\}} \prod_{\text{tensors } T} T_{s_v, a_1 a_2 \dots a_k}^{(v)}, \quad (45.68)$$

where s_v is the physical index at site v (dimension d) and a_i are bond indices (dimension χ , the *bond dimension*) connecting neighbouring tensors. The number of parameters scales as $N\chi^k d$ instead of d^N : polynomial rather than exponential.

The bond dimension χ controls the approximation quality: larger χ captures more entanglement. For $\chi = 1$: a product state (no entanglement). For $\chi \rightarrow d^{N/2}$: the exact state (maximum entanglement).

Matrix product states (MPS). White’s DMRG (1992), reformulated by Östlund and Rommer (1995) and by Verstraete and Cirac (2004), represents $|\Psi\rangle$ as a chain of matrices:

$$\Psi_{s_1 \dots s_N} = \text{Tr}(A^{s_1} A^{s_2} \dots A^{s_N}), \quad (45.69)$$

where A^{s_v} is a $\chi \times \chi$ matrix for each value of the physical index s_v . MPS are the natural ansatz for one-dimensional systems: they capture ground states of gapped Hamiltonians with area-law entanglement ($S_A \propto \text{const}$ for a contiguous block A).

MPS is the workhorse of modern condensed matter: it computes ground states, excitation spectra, real-time evolution, and finite-temperature properties of one-dimensional quantum systems with unprecedented accuracy — all within the Hamiltonian framework.

PEPS. Projected entangled pair states (Verstraete and Cirac, 2004) generalise MPS to two (and higher) dimensions: the tensors sit on the sites of a 2D lattice, with bond indices contracted along the lattice edges. PEPS capture area-law entanglement in 2D ($S_A \propto |\partial A|$, the boundary length of the region A).

MERA and the emergence of geometry

Vidal’s Multi-scale Entanglement Renormalization Ansatz (2007) is a tensor network with a specific hierarchical structure designed to capture entanglement at all length scales.

The structure. MERA consists of alternating layers of two types of tensors:

Disentanglers: unitary gates that remove short-range entanglement between neighbouring sites.

Isometries: maps that coarse-grain pairs of sites into single effective sites (reducing the number of sites by a factor of 2 at each layer).

The layers are stacked: the output of one layer (the coarse-grained sites) becomes the input of the next. The result is a tree-like network with $\sim \log_2 N$ layers, from the UV (the physical lattice at the bottom) to the IR (a single effective site at the top).

The geometry. Swingle (2012) observed that the structure of MERA is identical to the geometry of a time-slice of anti-de Sitter space:

The layers of MERA (from the UV at the bottom to the IR at the top) correspond to *radial slices of AdS* (from the boundary at $z = 0$ to the interior at $z = \infty$).

The number of sites at each layer decreases by a factor of 2, just as the proper area of a surface in AdS decreases with the radial coordinate z : $\text{Area} \propto (L/z)^{d-1}$.

The bond dimension χ at a given layer controls the amount of entanglement that can be described at that scale. In AdS: the corresponding quantity is the cross-sectional area of the bulk in Planck units: $\chi \sim e^{A/(4G)}$.

The top tensor (the single effective site at the top of the MERA tree) corresponds to the *deep interior* of AdS (the region $z \rightarrow \infty$, near the Poincaré horizon).

The MERA tensor network does not use the AdS metric as an input. It uses only the Hamiltonian of the boundary system (the lattice Hamiltonian H) and the variational optimisation of the tensor network (minimising $\langle \Psi | H | \Psi \rangle$). The AdS geometry emerges from the pattern of entanglement in the optimised state — it is read off, not put in.

Ryu–Takayanagi from tensor networks

The Ryu–Takayanagi formula (section 45.14):

$$S_A = \frac{\text{Area}(\gamma_A)}{4G_N} \quad (45.70)$$

(the entanglement entropy of a boundary region A equals the area of the minimal surface γ_A in the bulk that is homologous to A) has a direct tensor-network interpretation.

In a tensor network: the entanglement entropy of region A is bounded by the number of bonds cut by any surface separating A from its complement:

$$S_A \leq |\gamma| \cdot \ln \chi, \quad (45.71)$$

where $|\gamma|$ is the number of cut bonds (the minimal cut) and χ is the bond dimension. The minimal cut in the tensor network IS the minimal surface in AdS. The bond dimension χ plays the role of $e^{1/(4G)}$: it sets the “area per bond” in Planck units.

This is the most concrete realisation of the engine-rendering architecture for gravity:

The engine: a Hamiltonian H on a lattice, with ground state $|\Psi\rangle$ represented as a MERA tensor network.

The geometry: the AdS metric, read off from the entanglement structure of $|\Psi\rangle$ (the pattern of bond dimensions and minimal cuts in the MERA).

The Ryu–Takayanagi formula: a statement about the entanglement entropy of the tensor network, which happens to equal the area of a surface in the emergent geometry.

No metric is input. No graviton is invoked. No Einstein equation is postulated. The geometry emerges from the entanglement of the Fock-layer state.

Quantum error correction and the bulk

Almheiri, Dong, and Harlow (2015) showed that the AdS/CFT dictionary has the structure of a *quantum error-correcting code*: a bulk operator $\phi(x)$ deep in AdS is encoded redundantly in the boundary state, and the encoding is robust against erasure of a portion of the boundary.

In tensor-network language: a bulk site in the interior of the MERA is connected to many boundary sites through the network of tensors. The information at the bulk site is distributed (encoded) across the boundary. Erasing a small portion of the boundary (tracing out a subregion) does not destroy the bulk information — it can be reconstructed from the remaining boundary, just as a quantum error-correcting code protects information against partial erasure.

This gives a physical interpretation of the holographic dictionary: the bulk spacetime is the *code space* of a quantum error-correcting code, and the boundary state is the *physical encoding*. The “extra dimension” (the radial direction of AdS) is the depth of the encoding: deeper in the bulk means more redundantly encoded, more robust against boundary errors.

The Fock-Born reading

Tensor networks make the engine-rendering architecture for gravity explicit and computable:

Concept	Tensor network	AdS/CFT
Engine	Lattice Hamiltonian H	Boundary CFT
State	$ \Psi\rangle$ (MERA)	CFT ground state
Geometry	Pattern of bonds and cuts	AdS metric $g_{\mu\nu}$
Radial direction	Layer depth in MERA	Holographic z -coordinate
Entanglement entropy	Minimal cut $\times \ln \chi$	$\text{Area}(\gamma_A)/(4G)$
Bulk operator	Interior tensor	Local field $\phi(x)$
Robustness	Quantum error correction	Subregion duality
Newton’s constant	$1/\ln \chi$	$1/(4G)$ per Planck area

The tensor network is the engine object. The AdS geometry is the rendering. The Ryu–Takayanagi formula is the bridge: it translates the engine’s entanglement (a property of the Fock-layer state $|\Psi\rangle$) into the rendering’s geometry (the area of a surface in the Born-layer spacetime).

The tensor network does not invoke strings, gravitons, or the Einstein equation. It uses only a lattice Hamiltonian, a variational ansatz, and the entanglement structure of the ground state. The geometry emerges. The engine is non-gravitational. The rendering is gravitational. This is the architecture of section 45.34, realised concretely.

45.16 First and second quantisation of strings

The formulation of string theory in sections 45.7–45.8 is *first-quantised*: the string is treated as a single quantum-mechanical object (the worldsheet), and scattering amplitudes are computed by summing over worldsheet topologies. This is analogous to the worldline formalism for point particles, where the particle traces a 1D path in spacetime and loop diagrams are sums over worldline topologies (circles, figure-eights, etc.).

For point particles, first quantisation has well-known limitations that are resolved by *second quantisation*: promoting the particle coordinates to fields $\phi(x)$ on spacetime, with a Fock space of multi-particle states and a field-theoretic action $S[\phi]$. The passage from first to second quantisation is the passage from quantum mechanics to quantum field theory.

For strings: the analogous passage is from the worldsheet formulation to *string field theory* (SFT), where the string configuration $X^\mu(\sigma)$ is promoted to a field $\Phi[X(\sigma)]$ — a functional on the space of all string configurations. This passage is technically far more difficult than for point particles, and it remains incomplete in several respects.

First quantisation: the worldsheet

The framework. A single string propagates in a background spacetime $(G_{\mu\nu}, B_{\mu\nu}, \Phi)$. The worldsheet Σ (parametrised by (τ, σ)) is a 2D surface embedded in the target space. The worldsheet fields $X^\mu(\tau, \sigma)$ (the embedding coordinates) and their fermionic partners (in the superstring) are quantised as a 2D conformal field theory.

The Hilbert space of the worldsheet CFT contains all single-string states: the vacuum $|0; p\rangle$ (a string with centre-of-mass momentum p), the first excited state $\alpha_{-1}^\mu |0; p\rangle$ (a massless vector), the graviton $\epsilon_{\mu\nu} \alpha_{-1}^\mu \bar{\alpha}_{-1}^\nu |0; p\rangle$, and the infinite tower of massive modes.

Scattering amplitudes. Multi-string scattering is computed by inserting *vertex operators* on the worldsheet:

$$\mathcal{A}_n = \sum_{g=0}^{\infty} g_s^{2g-2+n} \int_{\mathcal{M}_{g,n}} d^d g \tau \left\langle \prod_{i=1}^n V_i(k_i, z_i) \right\rangle_{\Sigma_g}. \quad (45.72)$$

Each vertex operator $V_i(k_i, z_i)$ represents an external string state with momentum k_i , inserted at position z_i on the worldsheet. The integral over z_i (and over the moduli τ of the Riemann surface Σ_g) sums over all possible insertion points and worldsheet geometries.

Vertex operators as partial second quantisation. The vertex operator does something remarkable: it creates (or annihilates) a string quantum at a point on the worldsheet. For the graviton, the vertex operator is:

$$V_{\text{graviton}}(k, z, \bar{z}) = \epsilon_{\mu\nu} \partial X^\mu(z) \bar{\partial} X^\nu(\bar{z}) e^{ik \cdot X(z, \bar{z})}, \quad (45.73)$$

where ∂X^μ is the holomorphic worldsheet current and $e^{ik \cdot X}$ is the plane-wave factor. Inserting this operator on the worldsheet is equivalent to adding an external graviton to the scattering process.

In a point-particle QFT: external states are added by LSZ reduction of the field $\phi(x)$ — an operation that belongs to the second-quantised framework. In string theory: the vertex operator performs the same function *within* the first-quantised framework, without ever introducing a string field. The operator product expansion (OPE) of vertex operators encodes the three-point couplings, the fusion rules, and the factorisation properties that would normally require a field-theoretic action.

This is why the first-quantised string is so powerful: the vertex operator algebra contains much of the information that, for point particles, requires the full apparatus of QFT.

The limits of first quantisation

Despite the power of vertex operators, the first-quantised formulation has fundamental limitations — the same limitations that motivate the passage from quantum mechanics to QFT for point particles.

1. On-shell only. Vertex operators satisfy the physical state conditions ($L_0 |\text{phys}\rangle = 0$, etc.): they create *on-shell* string states. The scattering amplitudes (45.72) are on-shell S -matrix elements. There is no natural way to compute *off-shell* quantities — Green’s functions at arbitrary momenta, effective actions, or correlation functions of composite operators — within the first-quantised framework.

In QFT: off-shell quantities are essential for renormalisation, for computing effective potentials, for matching to low-energy effective theories, and for studying symmetry breaking. The first-quantised string cannot do any of this directly.

2. Perturbative only. The genus expansion (45.72) is a sum over worldsheet topologies, ordered by the string coupling g_s . Each term (each genus g) is a well-defined integral, but the sum $\sum_g g_s^{2g}(\dots)$ is an asymptotic series: it diverges at any finite g_s .

Non-perturbative effects — D-brane instantons ($\sim e^{-1/g_s}$), tunnelling between vacua, topology change — are invisible in the genus expansion. They were discovered not from the worldsheet but from consistency arguments, dualities, and the second-quantised formulation.

3. Background dependence. The worldsheet CFT is defined on a specific background ($G_{\mu\nu}, B_{\mu\nu}, \Phi$). Different backgrounds define different CFTs, with different vertex operators and different scattering amplitudes. The first-quantised formulation cannot describe a transition between backgrounds, because the background is an input, not an output.

In QFT: the background is a specific field configuration (a solution of the classical equations of motion), and the field $\phi(x)$ fluctuates around it. The field can tunnel between different backgrounds (different vacua). The first-quantised string has no analogue of this: the worldsheet CFT is stuck on its background.

4. No vacuum dynamics. The first-quantised formulation cannot describe the selection of the vacuum, the cosmological evolution of the background, or the real-time formation and evaporation of black holes. These require a dynamical framework in which the background itself is a degree of freedom — which is what string field theory aims to provide.

Second quantisation: string field theory

String field theory (SFT) promotes the string to a field. The fundamental variable is the *string field*:

$$|\Phi\rangle = \int \mathcal{D}X \, \Phi[X(\sigma)] |X(\sigma)\rangle, \quad (45.74)$$

where $\Phi[X(\sigma)]$ is a functional of the string configuration $X^\mu(\sigma)$ (the position of every point on the string at a fixed time), and $|X(\sigma)\rangle$ is the corresponding state in the first-quantised Hilbert space.

The string field $|\Phi\rangle$ is an element of the first-quantised Hilbert space $\mathcal{H}_{1\text{st}}$, but it is now treated as a *classical variable* in the action $S[\Phi]$, just as the classical field $\phi(x)$ in QFT is an element of the single-particle Hilbert space (a function on spacetime) treated as a classical variable.

Expanding $|\Phi\rangle$ in the oscillator basis of $\mathcal{H}_{1\text{st}}$:

$$|\Phi\rangle = \int d^d p \, [\phi(p) |0; p\rangle + A_\mu(p) \alpha_{-1}^\mu |0; p\rangle + h_{\mu\nu}(p) \alpha_{-1}^\mu \bar{\alpha}_{-1}^\nu |0; p\rangle + \dots]. \quad (45.75)$$

The string field contains *all* component fields simultaneously: a scalar $\phi(p)$ (the tachyon, absent in the superstring), a vector $A_\mu(p)$ (the gauge boson), a symmetric tensor $h_{\mu\nu}(p)$ (the graviton), and the infinite tower of massive modes. String field theory is a field theory with infinitely many component fields, all packaged into one functional $\Phi[X(\sigma)]$.

Witten's open string field theory

The most explicit construction (Witten 1986) is for the open bosonic string. The action:

$$S[\Phi] = -\frac{1}{2}\langle\Phi|Q_B|\Phi\rangle - \frac{g_o}{3}\langle\Phi|\Phi*\Phi\rangle, \quad (45.76)$$

where:

Q_B is the *BRST operator* of the first-quantised string. It encodes the free string dynamics (the worldsheet equation of motion, the physical state conditions, and the gauge invariance). The kinetic term $\langle\Phi|Q_B|\Phi\rangle$ is the string analogue of $\int \phi(-\nabla^2 + m^2)\phi$: it propagates each component field with the correct mass and spin.

$*$ is the *star product*: a non-commutative, associative multiplication on string fields. The star product $\Phi*\Psi$ glues the right half of string Φ to the left half of string Ψ , producing a new string. Geometrically: two open strings join at their midpoints. The cubic vertex $\langle\Phi|\Phi*\Phi\rangle$ is the string analogue of the ϕ^3 coupling: three strings join at a common midpoint.

g_o is the open string coupling constant.

The action (45.76) is a Chern–Simons-like theory: $S = \langle\Phi|Q_B|\Phi\rangle + \frac{2g_o}{3}\langle\Phi|\Phi*\Phi\rangle$ has the same algebraic structure as the Chern–Simons action $S_{CS} = \int \text{Tr}(A \wedge dA + \frac{2}{3}A \wedge A \wedge A)$, with Q_B playing the role of the exterior derivative d and $*$ playing the role of the wedge product $\wedge$.

Gauge invariance. The action is invariant under:

$$\delta|\Phi\rangle = Q_B|\Lambda\rangle + g_o(\Phi*\Lambda - \Lambda*\Phi), \quad (45.77)$$

where $|\Lambda\rangle$ is a string field gauge parameter. This contains all the gauge transformations of the component fields: the $U(1)$ (or $U(N)$) gauge transformation of A_μ , the diffeomorphism of $h_{\mu\nu}$, and the higher-spin gauge transformations of the massive modes — all packaged into a single string field transformation.

Technical difficulties

Both quantisation schemes have deep technical obstacles.

Difficulties of first quantisation.

Supermoduli: for the superstring, the moduli space $\mathcal{M}_{g,n}$ is replaced by a *supermoduli space* $\mathfrak{M}_{g,n}$ — a supermanifold with both bosonic and fermionic coordinates. The integration theory on supermanifolds is subtle: the measure is not a simple differential form but a Berezinian, and the choice of integration cycle (the “picture” for the vertex operators) introduces ambiguities. The proof that the amplitudes are well-defined and independent of these choices is complete for genus $g \leq 2$ but remains open at higher genus.

Divergence of the genus expansion: the sum $\sum_g g_s^{2g} \mathcal{A}_n^{(g)}$ is an asymptotic series with zero radius of convergence. The coefficients grow as $\mathcal{A}_n^{(g)} \sim (2g)!$ (factorial growth), similar to the perturbative expansion in QFT. The series must be Borel-resummed or supplemented by non-perturbative contributions ($\sim e^{-c/g_s}$) to define the full amplitude.

Measure ambiguities at high genus: the integration over $\mathcal{M}_{g,n}$ for $g \geq 3$ involves delicate cancellations between different spin structures (boundary components of the moduli space), and the proper definition of the measure requires the full machinery of algebraic geometry (the Deligne–Mumford compactification, the Mumford form).

Difficulties of second quantisation.

Infinite field content: the string field contains infinitely many component fields (one per oscillator level). The action (45.76), when expanded in components, generates an infinite number of interaction vertices with an infinite number of derivatives — a non-local field theory. Standard QFT tools (renormalisation, effective actions, lattice regularisation) are designed for theories with finitely many fields and break down or require major modification.

The star product: the star product $\Phi * \Psi$ involves gluing the string at its midpoint, which is a non-local operation in target space (it involves all Fourier modes of the string simultaneously). Computing Feynman diagrams in SFT requires evaluating star products of oscillator states — technically demanding and computationally expensive.

Closed strings: Witten’s cubic action works for open strings. For closed strings: the interaction is not cubic but involves vertices of all orders (3-point, 4-point, 5-point, $\dots$), each with a specific geometric structure (Zwiebach 1993). The closed SFT action: $S = \sum_{n=2}^{\infty} \frac{g_c^{n-2}}{n!} \{\Phi^n\}$, where $\{\Phi^n\}$ are the n -string vertices defined by specific regions of $\mathcal{M}_{0,n}$. The construction is exact but unwieldy: infinitely many vertices, each involving an integral over a portion of moduli space.

Superstring field theory: constructing a covariant SFT action for the superstring (Type II or heterotic) is the most technically challenging problem. The difficulties include: the picture-changing operators (which convert between different “pictures” of the superconformal ghosts), the Ramond sector (which requires special treatment due to the zero modes of the superghosts), and the constraint algebra (which is more complicated than for the bosonic string). Complete constructions exist (Berkovits 1995, Erler–Konopka–Sachs 2014, Sen 2015) but are technically demanding and not yet fully developed computationally.

Level truncation: in practice, SFT calculations truncate the string field to a finite number of component fields (the lowest N oscillator levels). The results converge as N increases, but the convergence is not rigorously established and the truncation breaks the gauge invariance at each level.

Successes of string field theory.

Despite the difficulties, SFT has achieved results inaccessible to the first-quantised formulation:

Tachyon condensation: Sen (1999) conjectured that the open bosonic string tachyon condenses to a new vacuum in which the D-brane on which the string ends has annihilated. The energy difference between the false vacuum (D-brane present) and the true vacuum (D-brane absent) equals the D-brane tension T_p . Schnabl (2005) found the exact analytic solution of Witten’s SFT describing the tachyon vacuum, confirming Sen’s conjecture.

Background independence: in SFT, different backgrounds correspond to different solutions of the equation of motion $Q_B \Phi + g_o \Phi * \Phi = 0$. The theory can interpolate between backgrounds by following the classical solution — something the first-quantised formulation cannot do.

The Fock-Born comparison

	First quantised (worldsheet)	Second quantised (SFT)
Fundamental object	One string on Σ	String field $\Phi[X(\sigma)]$
Hilbert space	$\mathcal{H}_{1\text{st}}$: single-string states	$\mathcal{F}_{\text{string}}$: multi-string Fock space
Interaction	Topology change of Σ	Cubic vertex $\Phi * \Phi$
External states	Vertex operators V_i	Asymptotic string field states
Off-shell?	No (vertex ops. are on-shell)	Yes (action is off-shell)
Non-perturbative?	No (genus expansion only)	In principle (classical solutions)
Background	Fixed (input to CFT)	Dynamical (solution of EOM)
Point-particle analogue	Worldline QM	Quantum field theory

The first-quantised string is to string field theory as quantum mechanics is to quantum field theory: a single-particle description that captures the on-shell, perturbative physics but misses the off-shell, non-perturbative, and vacuum-dynamical aspects that require the full Fock space.

The vertex operator is the mechanism by which the first-quantised framework borrows from the second-quantised one: inserting a vertex operator on the worldsheet is the first-quantised analogue of acting with a creation operator on the Fock vacuum. The vertex operator algebra (the OPE) encodes the same information as the cubic vertex $\Phi * \Phi$ of the string field action. This is why the first-quantised string is so much more powerful than first-quantised point-particle QM: the 2D conformal structure of the worldsheet provides algebraic access to information that, for point particles, requires the full field-theoretic machinery.

But the access is partial. The vertex operator creates on-shell states only; the string field creates arbitrary (on- and off-shell) states. The genus expansion sums perturbative topologies only; the string field action describes non-perturbative vacua. The worldsheet CFT is defined on a fixed background; the string field action is (in principle) background-independent.

The passage from first to second quantisation for strings remains one of the central technical challenges of the subject. For the open bosonic string: the construction is complete (Witten's cubic SFT). For the closed bosonic string: the construction exists but is unwieldy (Zwiebach's closed SFT). For the superstring: constructions exist but are technically demanding and computationally underdeveloped. A fully non-perturbative, background-independent formulation of string theory — the string-theoretic analogue of lattice QFT — does not yet exist.

45.17 Compactification: from ten dimensions to four

String theory is consistent in ten spacetime dimensions (or eleven, for M-theory). The observed world has four. The six (or seven) extra dimensions must be hidden — compactified to a size too small to be resolved at accessible energies. The geometry of the compact space determines all the low-energy physics: the gauge group, the matter content, the number of generations, and the values of the coupling constants.

Kaluza–Klein compactification: the mechanism

The idea (Kaluza 1921, Klein 1926): a higher-dimensional theory, when one or more dimensions are compactified to a small size, produces a lower-dimensional theory with additional fields.

The simplest case: 5D gravity on $M_4 \times S^1$. Consider pure gravity in $D = 5$ dimensions,

with the fifth dimension y compactified on a circle of radius R : $y \sim y + 2\pi R$. The 5D metric g_{MN} ($M, N = 0, \dots, 4$) decomposes into 4D fields:

$$g_{MN}(x, y) \rightarrow \begin{cases} g_{\mu\nu}(x) & \text{the 4D metric (gravity)} \\ g_{\mu 5}(x) = A_\mu(x) & \text{a 4D vector (electromagnetism)} \\ g_{55}(x) = e^{2\sigma(x)} & \text{a 4D scalar (the radion)} \end{cases} \quad (45.78)$$

The 5D Einstein equation $R_{MN}^{(5)} = 0$ reduces to the 4D Einstein equation plus the Maxwell equation for A_μ plus a scalar equation for σ . Gravity in five dimensions *contains* electromagnetism in four dimensions: the gauge field is the metric component along the compact direction.

The Kaluza–Klein tower. A field $\Phi(x, y)$ in 5D, periodic in y , expands in Fourier modes:

$$\Phi(x, y) = \sum_{n=-\infty}^{\infty} \phi_n(x) e^{iny/R}. \quad (45.79)$$

The 5D wave equation $(\square_4 + \partial_y^2)\Phi = 0$ gives a 4D equation for each mode: $(\square_4 - n^2/R^2)\phi_n = 0$. Each mode ϕ_n is a 4D field with mass $m_n = |n|/R$.

The $n = 0$ mode is massless: it is the zero mode, the field visible at energies $E \ll 1/R$. The $n \neq 0$ modes are massive: they form the *KK tower*, a ladder of states with masses $1/R, 2/R, 3/R, \dots$. At energies $E \ll 1/R$: only the zero modes are excited, and the physics is effectively four-dimensional. The extra dimension is “invisible” — it manifests only through the KK tower at $E \sim 1/R$.

The general case. For a D -dimensional theory on $M_4 \times K$ (where K is a compact $(D - 4)$ -dimensional manifold):

The 4D massless fields are determined by the *harmonic forms* on K (the zero modes of the Laplacian ∇_K^2 on K). Scalars come from harmonic 0-forms; vectors from harmonic 1-forms; 2-forms from harmonic 2-forms; etc. The number of each type is a topological invariant of K : the Betti numbers $b_p = \dim H^p(K)$.

The 4D massive fields form KK towers with masses $m \sim 1/L$ (where L is the characteristic size of K) and represent the “excitation modes” of the compact space. If $L \sim \ell_P$: the KK masses are $\sim M_P$ and the tower is unobservable.

What string theory requires of the compact space

The 10D superstring theories require specific properties of the compact manifold K_6 (a 6-dimensional manifold) to preserve desirable features of the 4D physics:

$\mathcal{N} = 1$ *supersymmetry in 4D* (phenomenologically motivated: $\mathcal{N} = 1$ SUSY is the minimal amount compatible with chiral fermions, and it stabilises the hierarchy — if it exists at the TeV scale).

The condition for $\mathcal{N} = 1$ SUSY in 4D, starting from $\mathcal{N} = 2$ in 10D (Type II) or $\mathcal{N} = 1$ in 10D (heterotic): the compact manifold K_6 must admit a covariantly constant spinor η :

$$\nabla_m \eta = 0, \quad m = 1, \dots, 6. \quad (45.80)$$

This condition is equivalent to:

(i) K_6 is a *Kähler manifold*: a complex manifold with a Hermitian metric compatible with the complex structure.

(ii) The *first Chern class vanishes*: $c_1(K_6) = 0$. By Yau's theorem (1978, proving the Calabi conjecture of 1957): this is equivalent to the existence of a Ricci-flat Kähler metric $R_{mn} = 0$ on K_6 .

(iii) The holonomy group is $SU(3) \subset SO(6)$: the parallel transport of a spinor around any closed loop returns it to itself (reduced holonomy).

A compact Kähler manifold with $c_1 = 0$ is a *Calabi–Yau threefold* (“threefold” because the complex dimension is 3, the real dimension is 6).

The topology of Calabi–Yau manifolds

The topology of a Calabi–Yau threefold is characterised by its *Hodge numbers* $h^{p,q}$ (the dimensions of the Dolbeault cohomology groups $H^{p,q}(K_6)$). For a CY threefold: $h^{0,0} = h^{3,3} = 1$ (connected), $h^{1,0} = h^{0,1} = 0$ (simply connected), and the non-trivial Hodge numbers are $h^{1,1}$ and $h^{2,1}$.

$h^{1,1}$: the Kähler moduli. $h^{1,1}$ counts the harmonic $(1,1)$ -forms on K_6 . Each corresponds to a deformation of the Kähler form J (the metric’s “size” parameters): the volumes of the 2-cycles, 4-cycles, and the overall volume of K_6 . In the 4D theory: each Kähler modulus is a massless scalar field.

$h^{2,1}$: the complex structure moduli. $h^{2,1}$ counts the harmonic $(2,1)$ -forms on K_6 . Each corresponds to a deformation of the complex structure (the “shape” parameters): how the 3-cycles are embedded. In the 4D theory: each complex structure modulus is a massless scalar field.

The Euler characteristic and the number of generations. For the heterotic string on a CY threefold with a vector bundle V : the number of chiral fermion generations in 4D is:

$$N_{\text{gen}} = \frac{|\chi(K_6)|}{2}, \quad \chi = 2(h^{1,1} - h^{2,1}), \quad (45.81)$$

where χ is the Euler characteristic. Three generations ($N_{\text{gen}} = 3$) requires $|\chi| = 6$: a specific topological constraint on the CY manifold. This is one of the most striking predictions of string compactification: the number of fermion generations is determined by the topology of the extra dimensions.

The simplest example: the quintic threefold. The zero set of a degree-5 polynomial in $\mathbb{CP}^4$: $\{z \in \mathbb{CP}^4 : p_5(z_0, \dots, z_4) = 0\}$. The Hodge numbers: $h^{1,1} = 1$, $h^{2,1} = 101$, $\chi = -200$. This gives $N_{\text{gen}} = 100$ — far too many. But more complicated CY manifolds (complete intersections, hypersurfaces in toric varieties, quotients by discrete groups) can have $|\chi| = 6$ and hence three generations.

The moduli problem and flux compactifications

At tree level in string perturbation theory: the Kähler and complex structure moduli are exactly massless. Their values are free parameters — they parametrise a continuous family of vacua (the moduli space). Each point in moduli space gives a different 4D physics (different gauge couplings, different Yukawa couplings, different mass ratios).

This is the *moduli problem*: the compactification has flat directions — massless scalar fields with no potential. These would mediate long-range forces (fifth forces) and would have been detected. The moduli must acquire masses.

Flux compactifications. Turn on quantised fluxes $\int_{\Sigma} F_p = n \in \mathbb{Z}$ of the p -form field strengths through the non-trivial cycles Σ of the CY manifold. The fluxes generate a potential $V(\phi_i)$ for

the moduli ϕ_i through the Gukov–Vafa–Witten superpotential:

$$W = \int_{K_6} \Omega \wedge (F_3 - \tau H_3), \quad (45.82)$$

where Ω is the holomorphic $(3,0)$ -form, F_3 and H_3 are the RR and NS 3-form fluxes, and $\tau = C_0 + ie^{-\Phi}$ is the axion-dilaton. The superpotential stabilises the complex structure moduli and the dilaton at $D_i W = 0$ (supersymmetric minima).

The Kähler moduli are stabilised by non-perturbative effects (D-brane instantons or gaugino condensation on D7-branes) in the KKLT construction (Kachru, Kallosh, Linde, Trivedi, 2003). An anti-D3-brane at the tip of a warped throat breaks SUSY and lifts the vacuum energy to a (metastable) de Sitter minimum.

Why Calabi–Yau is absent from AdS/CFT

The original and most-studied example of AdS/CFT (section 45.13) is Type IIB on $\text{AdS}_5 \times S^5$. The internal space is S^5 , not a Calabi–Yau manifold. The reasons are both historical and structural:

S^5 preserves maximal supersymmetry. S^5 has $SO(6) \cong SU(4)$ isometry, which becomes the $SU(4)_R$ R-symmetry of $\mathcal{N} = 4$ SYM in 4D. Maximal SUSY (16 Poincaré + 16 conformal supercharges) makes the dual CFT exactly conformal (the beta function vanishes to all orders), exactly solvable in the planar limit, and amenable to integrability methods. The theory is a laboratory, not a model of the real world.

Calabi–Yau reduces supersymmetry. Replacing S^5 by a Sasak–Einstein manifold Y_5 (whose cone is a CY 3-fold) reduces the SUSY to $\mathcal{N} = 1$ in 4D. The dual CFT is a less supersymmetric, less tractable gauge theory (a “quiver” gauge theory determined by the singularity structure of the CY cone). These AdS/CFT examples exist ($\text{AdS}_5 \times T^{1,1}$, the Klebanov–Witten theory; $\text{AdS}_5 \times Y^{p,q}$) but are technically harder and less fully understood.

AdS $\neq$ the real world. AdS/CFT gives a spacetime with a negative cosmological constant ($\Lambda < 0$). CY compactifications aim for Minkowski ($\Lambda = 0$) or de Sitter ($\Lambda > 0$). The phenomenological programme (CY compactification with fluxes, KKLT, etc.) and the AdS/CFT programme (exact duality in a controlled setting) have different goals:

CY compactifications ask: can string theory reproduce the Standard Model (gauge group, matter content, three generations, coupling constants) in 4D? This requires a specific CY manifold with specific fluxes and brane configurations.

AdS/CFT asks: can string theory provide exact answers about strongly coupled gauge theories? This requires maximal symmetry and computational control, which S^5 provides and CY does not.

The two programmes are complementary. They use different internal geometries because they ask different questions. CY compactification is about phenomenology (the real world); AdS/CFT is about exact results (a theoretical laboratory). Neither renders the other obsolete.

Mirror symmetry

For every Calabi–Yau manifold K with Hodge numbers $(h^{1,1}, h^{2,1})$, there exists (conjecturally, and proven in many cases) a *mirror manifold* $\tilde{K}$ with Hodge numbers $(h^{2,1}, h^{1,1})$: the two Hodge numbers are exchanged.

The mirror symmetry conjecture (Dixon, Lerche, Vafa, Warner, 1989):

$$\text{Type IIA on } K \longleftrightarrow \text{Type IIB on } \tilde{K}. \quad (45.83)$$

The two compactifications give the same 4D physics. The Kähler moduli of K (controlling the sizes of 2-cycles) are mapped to the complex structure moduli of $\tilde{K}$ (controlling the shapes of 3-cycles), and vice versa.

Why mirror symmetry is powerful. Some calculations are easy on one side and hard on the other. The most celebrated example:

Counting rational curves on K (the Gromov–Witten invariants): on the IIA side, this counts the instantons (holomorphic maps from $\mathbb{CP}^1$ into K). This is a hard enumerative geometry problem.

Period integrals on $\tilde{K}$: on the IIB side, the same information is encoded in the periods $\Pi_i = \int_{\gamma_i} \Omega$ of the holomorphic 3-form Ω over the 3-cycles γ_i . This is a classical differential equation (the Picard–Fuchs equation).

Candelas, de la Ossa, Green, and Parkes (1991) used mirror symmetry to compute the number of rational curves of each degree on the quintic threefold — solving a problem that had been open in algebraic geometry for over a century. The result (for degree d): $n_1 = 2875$, $n_2 = 609250$, $n_3 = 317206375$, $\dots$, obtained by solving a single differential equation on the mirror.

Mirror symmetry is the deepest geometric consequence of the web of dualities: it exchanges “size” with “shape,” turning hard counting problems into easy calculus problems.

The landscape

The number of distinct Calabi–Yau manifolds is large but finite (the known constructions give $\sim 10^5$ topologically distinct CY threefolds). But the number of *vacua* is much larger, because each CY can be equipped with different fluxes, brane configurations, and moduli stabilisation mechanisms.

The fluxes are quantised integers: $\int_{\Sigma_i} F_p = n_i \in \mathbb{Z}$. For a CY with $b_3 \sim 300$ independent 3-cycles (typical for the flux compactifications studied in detail): each flux integer n_i can range over ~ 10 values (bounded by the tadpole cancellation condition $\sum n_i^2 \leq L$). The number of flux choices: $\sim 10^{300}$.

Combined with the number of CY manifolds, brane configurations, and discrete choices (orientifold projections, Wilson lines): the estimated number of vacua is:

$$N_{\text{vac}} \sim 10^{500} \quad (\text{the Bousso–Polchinski estimate, 2000}). \quad (45.84)$$

Each vacuum has a different low-energy physics: a different gauge group, different matter content, different coupling constants, different cosmological constant. This is the *string theory landscape*: an enormous ensemble of consistent 4D effective theories, all derived from the same 10D string theory.

The landscape problem. If string theory has 10^{500} vacua: it does not predict the gauge group, the matter content, the number of generations, or the cosmological constant. It predicts only that they take values somewhere in the landscape. The specific values observed in our universe are selected by initial conditions (or by anthropic reasoning), not by the theory.

This is either:

A failure of predictivity: string theory was supposed to predict a unique vacuum (the Standard Model with its specific parameters). The landscape shows it predicts 10^{500} vacua, of which ours is one. The theory is unfalsifiable (any observation is compatible with some vacuum).

An explanation of fine-tuning: the cosmological constant Λ takes $\sim 10^{500}$ different values across the landscape. Some of these values are small enough to allow structure formation. If we live in

a vacuum selected (by the anthropic principle) from the subset with small Λ : the smallness of Λ is explained without fine-tuning.

The swampland programme. Not every low-energy effective field theory is consistent with quantum gravity. The *swampland* is the set of EFTs that appear consistent at low energies but cannot be UV-completed in string theory. The swampland conjectures (Vafa 2005, Ooguri–Vafa 2007) propose criteria for distinguishing landscape (consistent) from swampland (inconsistent): for example, the weak gravity conjecture (gravity is always the weakest force), the distance conjecture (infinite-distance limits in moduli space produce infinite towers of light states), and the de Sitter conjecture (stable de Sitter vacua may not exist in string theory — which would rule out KKLT and the landscape of dS vacua).

The swampland programme turns the landscape problem into a research programme: instead of “string theory predicts anything,” it asks “what does string theory forbid?”

The Fock-Born reading

Compactification is a Born-layer operation: it specifies the geometry of the rendering.

The 10D string theory (the Fock-layer engine, to the extent that it is defined) is unique (or unified by M-theory). The 10^{500} vacua of the landscape are 10^{500} different Born-layer renderings of the same engine, obtained by choosing different CY manifolds, fluxes, and brane configurations. Each rendering produces a different 4D physics.

The landscape problem is then: the engine is defined, but the rendering is not unique. The engine does not select a preferred CY manifold or a preferred flux configuration. The rendering is an additional input, not an output.

In the lattice-first framework: the engine is a lattice Hamiltonian $H(a)$, and the rendering (the 4D physics) is computed from the lattice correlators. There is no compactification, no CY manifold, no flux choice. The “extra dimensions” and the “compact geometry” are properties of the rendering (the effective string description at $\lambda \gg a$), not of the engine. The lattice is four-dimensional; the apparent ten dimensions emerge in the rendered string theory as a low-energy effective description.

Whether the lattice-first framework can reproduce the CY- compactification phenomenology (gauge groups, matter content, three generations) without invoking extra dimensions is an open question. The CY manifold may be a useful organising principle for the Born-layer output, even if it does not correspond to a literal six-dimensional space in the Fock layer.

45.18 Moonshine: finite groups and modular forms

The deepest connection between string theory and pure mathematics is *moonshine*: a family of unexpected relationships between finite groups (discrete symmetries) and modular forms (functions with specific transformation properties under $SL(2, \mathbb{Z})$). String theory is not merely a tool for studying moonshine; it is the *explanation* of moonshine. The relationships that appear mysterious from the standpoint of number theory and group theory become natural — indeed inevitable — when viewed through the lens of the string worldsheet.

The j -function and the Monster

The j -function is the unique modular-invariant function on the upper half-plane $\mathbb{H}$ with a simple pole at the cusp:

$$j(\tau) = q^{-1} + 744 + 196884q + 21493760q^2 + 864299970q^3 + \cdots, \quad q = e^{2\pi i\tau}. \quad (45.85)$$

It is the generator of the field of modular functions: every $SL(2, \mathbb{Z})$ -invariant meromorphic function on $\mathbb{H}$ is a rational function of j .

The Monster group $\mathbb{M}$ is the largest sporadic simple group:

$$|\mathbb{M}| = 2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71 \approx 8 \times 10^{53}. \quad (45.86)$$

It has 194 conjugacy classes and 194 irreducible representations. The smallest non-trivial representation has dimension 196883.

McKay’s observation (1978). The coefficient 196884 in the j -function is $196884 = 196883 + 1$: the dimension of the smallest non-trivial representation of $\mathbb{M}$, plus the trivial representation. Thompson noticed that $21493760 = 21296876 + 196883 + 1$: the sum of the dimensions of the three smallest representations. And so on for higher coefficients.

This is not a coincidence. Every coefficient of $j(\tau) - 744$ is a sum of dimensions of irreducible representations of $\mathbb{M}$.

The Conway–Norton conjecture (1979). Conway and Norton conjectured that there exists a graded infinite-dimensional representation of $\mathbb{M}$:

$$V^{\natural} = \bigoplus_{n=-1}^{\infty} V_n, \quad \dim V_n = c_n \quad (\text{the } n\text{-th Fourier coefficient of } j - 744), \quad (45.87)$$

such that for every element $g \in \mathbb{M}$, the *McKay–Thompson series*:

$$T_g(\tau) = \sum_{n=-1}^{\infty} \text{Tr}(g|V_n) q^n \quad (45.88)$$

is a *hauptmodul*: the unique generator of the function field of a genus-zero quotient $\mathbb{H}/\Gamma_g$, where Γ_g is a specific discrete subgroup of $SL(2, \mathbb{R})$ depending on g . For $g = 1$ (the identity): $T_1(\tau) = j(\tau) - 744$.

This is *monstrous moonshine*: a precise, checkable conjecture relating every element of the Monster group to a specific modular function.

String theory enters: the moonshine module

Frenkel, Lepowsky, and Meurman (1988) constructed the moonshine module $V^{\natural}$ explicitly. The construction is a string theory:

Step 1: the Leech lattice. The Leech lattice Λ_{24} is the unique 24-dimensional even unimodular lattice with no vectors of norm 2 (no “short roots”). Its automorphism group is the Conway group Co_0 , of order $\sim 8 \times 10^{18}$.

Step 2: the bosonic string on Λ_{24} . Compactify the 26-dimensional bosonic string on $\mathbb{R}^{1,1} \times (\mathbb{R}^{24}/\Lambda_{24})$: the 24 transverse dimensions are compactified on the torus $\mathbb{R}^{24}/\Lambda_{24}$. The worldsheet CFT is a free bosonic theory with target space the Leech lattice torus. The partition function:

$$Z_{\Lambda_{24}}(\tau) = \frac{\Theta_{\Lambda_{24}}(\tau)}{\eta(\tau)^{24}}, \quad (45.89)$$

where $\Theta_{\Lambda_{24}}$ is the theta function of the Leech lattice and $\eta(\tau) = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n)$ is the Dedekind eta function.

Step 3: the $\mathbb{Z}_2$ orbifold. Orbifold the theory by the $\mathbb{Z}_2$ reflection $X^i \rightarrow -X^i$ ($i = 1, \dots, 24$): identify each lattice point with its negative. The orbifold introduces twisted sectors (strings

that close only up to the $\mathbb{Z}_2$ action) and projects out non-invariant states. The resulting vertex operator algebra (VOA) is $V^\natural$: the moonshine module.

Its partition function:

$$Z_{V^\natural}(\tau) = \text{Tr}_{V^\natural}(q^{L_0-1}) = j(\tau) - 744. \quad (45.90)$$

The j -function (minus 744) IS the partition function of the bosonic string on the Leech lattice orbifold.

The Monster as the symmetry group. The automorphism group of the vertex operator algebra $V^\natural$ is the Monster group $\mathbb{M}$. The Monster is the symmetry of the string theory — the group of transformations that preserve the worldsheet CFT.

The McKay–Thompson series $T_g(\tau) = \text{Tr}_{V^\natural}(g \cdot q^{L_0-1})$ are *twisted partition functions*: the partition function with a group element g inserted into the trace. In string theory: this is the partition function with a g -twisted boundary condition around the thermal circle (the orbifold by g). The genus-zero property (the hauptmodul condition) follows from specific properties of the VOA and was proven by Borchers (1992) using the “monster Lie algebra” and the no-ghost theorem of string theory.

Why string theory explains moonshine

The connection between the Monster and the j -function is explained by three features of string theory:

1. *Modular invariance.* The partition function of any consistent string theory is modular invariant: $Z(\tau) = Z(\tau+1) = Z(-1/\tau)$. This is a consequence of the worldsheet being a torus (for the one-loop vacuum amplitude): the modular group $SL(2, \mathbb{Z})$ is the group of reparametrisations of the torus that leave the physics invariant (section 45.8). The j -function is the unique modular-invariant function with a simple pole at the cusp — so any string compactification producing a modular-invariant partition function with the right pole structure must give j .

2. *The vertex operator algebra.* The worldsheet CFT of a string compactification is a vertex operator algebra: a mathematical structure encoding the operator product expansion, the state-operator correspondence, and the modular properties of the partition function. The VOA is the bridge between the physics (the string theory) and the mathematics (the modular forms and the finite group).

3. *Lattice symmetries.* The compactification lattice Λ_{24} has a large discrete symmetry group (Co_0) , which is extended by the $\mathbb{Z}_2$ orbifold construction to the full Monster group. The symmetry group of the lattice becomes the symmetry group of the string theory, which becomes the finite group whose representations organise the coefficients of the modular function.

Without string theory: the relationship between $\mathbb{M}$ and j is a mysterious numerical coincidence (“moonshine” — as in “distilled nonsense,” Conway’s original name). With string theory: it is the inevitable consequence of a specific compactification having a specific symmetry group.

Mathieu moonshine

Eguchi, Ooguri, and Tachikawa (2010) discovered a second moonshine phenomenon, involving a smaller sporadic group and a string-theoretic geometry.

The K3 surface. K3 is the unique compact Calabi–Yau 2-fold (complex dimension 2, real dimension 4): the simplest non-trivial CY manifold. It has Hodge numbers $h^{1,1} = 20$, $h^{2,0} = 1$, and Euler characteristic $\chi = 24$. Type IIA string theory compactified on K3 gives a 6D theory with $(1,1)$ supersymmetry.

The elliptic genus. The elliptic genus of K3:

$$Z_{K3}(\tau, z) = \text{Tr}_{\mathcal{H}_{RR}} [(-1)^{F_R} q^{L_0} \bar{q}^{\bar{L}_0} y^{J_0}], \quad y = e^{2\pi i z}, \quad (45.91)$$

is a topological invariant that counts BPS states (states preserved by some supercharges). It is a weak Jacobi form of weight 0 and index 1, and it has the expansion:

$$Z_{K3}(\tau, z) = 24 \mu(\tau, z) + \sum_{n \geq 0} A_n \text{ch}_n(\tau, z), \quad (45.92)$$

where μ is a specific Appell–Lerch sum and ch_n are $\mathcal{N} = 4$ superconformal characters.

The Mathieu group. The coefficients A_n are: $A_0 = -2$, $A_1 = 90$, $A_2 = 462$, $A_3 = 1540$, ... Eguchi, Ooguri, and Tachikawa observed: $90 = 45 + \overline{45}$, $462 = 231 + \overline{231}$, etc. — each A_n decomposes into dimensions of irreducible representations of the Mathieu group M_{24} (a sporadic simple group of order 244,823,040, the automorphism group of the binary Golay code).

This is *Mathieu moonshine*: the BPS degeneracies of the K3 sigma model are organised by representations of M_{24} , just as the coefficients of j are organised by representations of $\mathbb{M}$.

The physical interpretation: M_{24} acts as a symmetry on the BPS states of the K3 sigma model. Different K3 surfaces (different points in the moduli space) have different geometric symmetry groups, and no single K3 has M_{24} as its full symmetry. But the *combined* symmetries of all K3 surfaces, acting on the BPS states, generate M_{24} . The moonshine group is a “hidden symmetry” visible only in the BPS spectrum, not in the geometry of any individual K3.

Umbral moonshine

Cheng, Duncan, and Harvey (2012) vastly generalised the pattern.

In 24 dimensions, there are exactly 24 even unimodular lattices: the 23 *Niemeier lattices* (each with root vectors of norm 2, labelled by their root systems X) and the Leech lattice (no norm-2 vectors). Monstrous moonshine is associated to the Leech lattice.

For each Niemeier lattice X : there is a finite group G_X (related to the automorphism group of X) and a vector-valued mock modular form $H_X(\tau)$, such that the components of H_X are graded characters of G_X :

$$H_{X,r}(\tau) = \sum_n \text{Tr}(g|_{V_{n,r}^X}) q^n. \quad (45.93)$$

This is *umbral moonshine*: a moonshine relation for each of the 23 Niemeier lattices.

Mathieu moonshine is one case: the Niemeier lattice with root system A_1^{24} gives $G_X = M_{24}$ and H_X related to the K3 elliptic genus.

Umbral moonshine was conjectured by Cheng–Duncan–Harvey (2012) and proved by Duncan, Griffin, and Ono (2015). The proof uses Rademacher sums (exact formulas for the Fourier coefficients of mock modular forms) and does not rely on string theory directly, but the *existence* of the pattern is explained by string theory: each Niemeier lattice defines a string compactification, and the finite group G_X is (a subquotient of) the symmetry group of the corresponding CFT.

Further moonshines

The pattern continues to expand:

Thompson moonshine (Harvey and Rayhaun, 2015): the Thompson group Th (a sporadic group of order $\sim 9 \times 10^{16}$) and weight-1/2 modular forms.

O’Nan moonshine (Duncan, Mertens, Ono, 2017): the O’Nan group O’N (order $\sim 5 \times 10^{11}$) and weight-3/2 mock modular forms arising from Borchers products.

In each case: a sporadic finite group is connected to a specific class of modular or mock modular forms, and the connection is explained (or expected to be explained) by a string-theoretic or VOA construction.

The role of string theory: a summary

String theory explains moonshine at three levels:

Level 1: construction. The moonshine module $V^\natural$ is the Hilbert space of a specific string theory (the bosonic string on the Leech lattice orbifold). The Monster group is the symmetry group of this string theory. The j -function is its partition function. Without string theory: the connection is a numerical mystery. With string theory: it is a consequence of compactification on a specific lattice.

Level 2: modular invariance. The modular properties of the McKay–Thompson series (the hauptmodul/genus-zero condition) follow from the modular invariance of the string partition function. The $SL(2, \mathbb{Z})$ invariance of the worldsheet torus is the physical origin of the modular structure.

Level 3: organising principle. The vertex operator algebra provides the algebraic framework that unifies the finite group (the automorphism group of the VOA), the modular form (the partition function/character of the VOA), and the physical theory (the string compactification). The VOA is the mathematical distillation of the worldsheet CFT: it captures exactly the structure needed for moonshine.

Moonshine is not an application of string theory to a mathematical problem. It is evidence that the algebraic structures of string theory — vertex operator algebras, modular invariance, lattice compactifications — are deeply embedded in the structure of finite groups and number theory. The sporadic groups, which appear to be arbitrary accidents of the classification of finite simple groups, turn out to be symmetry groups of specific string compactifications. String theory reveals order in what appeared to be chaos.

Part II: Loop Quantum Gravity

45.19 Why go beyond the metric?

The Einstein–Hilbert action $S_{\text{EH}} = \frac{1}{16\pi G} \int d^4x \sqrt{-g} R$ is second-order in the metric: it contains R , which involves second derivatives of $g_{\mu\nu}$. To pass to a Hamiltonian formulation, one needs a first-order action — one with at most first derivatives — so that the Legendre transform to phase-space variables (q, p) is straightforward.

This is achieved by the *tetrad* (or *vierbein*) formalism: replace the metric $g_{\mu\nu}$ by a set of four orthonormal vector fields e_μ^I ($I = 0, 1, 2, 3$, the “internal” or “Lorentz” index) that define a local reference frame at each spacetime point. The metric is reconstructed from the tetrad:

$$g_{\mu\nu}(x) = e_\mu^I(x) e_\nu^J(x) \eta_{IJ}, \quad (45.94)$$

where $\eta_{IJ} = \text{diag}(-1, +1, +1, +1)$ is the Minkowski metric of the local frame.

The tetrad has 16 components (4×4); the metric has 10 (symmetric 4×4). The 6 extra components are the local Lorentz transformations: rotations and boosts of the local frame, $e_\mu^I \rightarrow \Lambda^I_J(x) e_\mu^J$, which leave $g_{\mu\nu}$ unchanged. This is a new gauge symmetry — the local Lorentz group $SO(3, 1)$ — that acts on the internal index I .

The tetrad formalism is essential for two reasons:

It couples gravity to fermions. Spinors transform under $SL(2, \mathbb{C})$, the double cover of $SO(3, 1)$, not under $GL(4, \mathbb{R})$ (the group of general coordinate transformations). To write a covariant

Dirac equation on a curved background, one needs a local frame e_μ^I to convert spacetime indices μ to Lorentz indices I . The tetrad is the soldering form that connects spacetime and the internal Lorentz frame.

It gives a first-order action. The tetrad e_μ^I and the spin connection ω_μ^{IJ} (below) are independent variables. The action is first-order in both, and the Hamiltonian formulation is direct.

45.20 The spin connection

The spin connection ω_μ^{IJ} is the gauge field of the local Lorentz group. It defines parallel transport of internal (Lorentz) indices along spacetime curves, just as the Christoffel connection $\Gamma_{\mu\nu}^\lambda$ defines parallel transport of spacetime indices.

The covariant derivative. For a field $V^I(x)$ carrying a Lorentz index:

$$D_\mu V^I = \partial_\mu V^I + \omega^I_{J\mu} V^J, \quad (45.95)$$

where $\omega_\mu^{IJ} = -\omega_\mu^{JI}$ (antisymmetric in I, J , because $SO(3,1)$ generators are antisymmetric). For a spinor ψ : $D_\mu \psi = \partial_\mu \psi + \frac{1}{4} \omega_\mu^{IJ} \gamma_I \gamma_J \psi$, where γ_I are the flat-space Dirac matrices.

The torsion. The torsion 2-form:

$$T_{\mu\nu}^I = D_\mu e_\nu^I - D_\nu e_\mu^I = \partial_\mu e_\nu^I - \partial_\nu e_\mu^I + \omega^I_{J\mu} e_\nu^J - \omega^I_{J\nu} e_\mu^J. \quad (45.96)$$

In GR without matter with spin: the torsion vanishes ($T^I = 0$), and this condition uniquely determines the spin connection in terms of the tetrad: $\omega_\mu^{IJ} = \omega_\mu^{IJ}(e)$ (the “compatible” or “torsion-free” connection). This is the Lorentz-frame analogue of the Levi-Civita connection: $\omega(e)$ is to e_μ^I as $\Gamma(g)$ is to $g_{\mu\nu}$.

The relation to the Christoffel connection. The tetrad postulate $D_\mu e_\nu^I \equiv \partial_\mu e_\nu^I + \omega^I_{J\mu} e_\nu^J - \Gamma_{\mu\nu}^\lambda e_\lambda^I = 0$ (the requirement that the total covariant derivative of e_μ^I vanishes) relates ω to Γ :

$$\omega^I_{J\mu} = e_\lambda^I e_J^\nu \Gamma_{\mu\nu}^\lambda + e_\lambda^I \partial_\mu e_J^\lambda. \quad (45.97)$$

The spin connection encodes the same geometric information as the Christoffel connection, expressed in the local Lorentz frame.

45.21 The curvature of the spin connection

The curvature 2-form of the spin connection:

$$F_{\mu\nu}^{IJ} = \partial_\mu \omega_\nu^{IJ} - \partial_\nu \omega_\mu^{IJ} + \omega^I_{K\mu} \omega_\nu^{KJ} - \omega^I_{K\nu} \omega_\mu^{KJ}. \quad (45.98)$$

This is the field strength of the $SO(3,1)$ gauge field ω_μ^{IJ} , exactly analogous to the Yang–Mills field strength $G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + gf^{abc} G_\mu^b G_\nu^c$ (section 29).

The curvature 2-form is related to the Riemann tensor by:

$$F_{\mu\nu}^{IJ} = e_\rho^I e_\sigma^J R_{\sigma\mu\nu}^\rho. \quad (45.99)$$

The Riemann tensor in the tetrad frame is the curvature of the spin connection. The Ricci scalar:

$$R = e_I^\mu e_J^\nu F_{\mu\nu}^{IJ}. \quad (45.100)$$

This is the key identity: it rewrites the Einstein–Hilbert Lagrangian $\sqrt{-g} R$ entirely in terms of e_μ^I and $F_{\mu\nu}^{IJ}(\omega)$, with no second derivatives of the metric.

45.22 The Palatini action

The first-order (Palatini) action treats e_μ^I and ω_μ^{IJ} as independent variables:

$$S_{\text{Pal}} = \frac{1}{16\pi G} \int d^4x |e| e_I^\mu e_J^\nu F_{\mu\nu}^{IJ}(\omega), \quad (45.101)$$

where $|e| = \det(e_\mu^I) = \sqrt{-g}$. This is first-order: it contains only first derivatives of ω (through F) and no derivatives of e .

Varying with respect to ω : the equation of motion forces the torsion to vanish, $T_{\mu\nu}^I = 0$, which determines $\omega = \omega(e)$ (the torsion-free connection).

Varying with respect to e : substituting $\omega = \omega(e)$ back into the equation of motion gives the Einstein equation $G_{\mu\nu} = 0$.

The Palatini action is classically equivalent to the Einstein–Hilbert action: same equations of motion, same physics. The advantage is that it is first-order and treats gravity as a gauge theory of the Lorentz group, with ω_μ^{IJ} as the gauge field and $F_{\mu\nu}^{IJ}$ as the field strength.

45.23 The Holst action and the Barbero–Immirzi parameter

Holst (1996) added a second term to the Palatini action:

$$S_{\text{Holst}} = \frac{1}{16\pi G} \int d^4x |e| e_I^\mu e_J^\nu \left(F_{\mu\nu}^{IJ} + \frac{1}{2\gamma} \varepsilon^{IJ}{}_{KL} F_{\mu\nu}^{KL} \right), \quad (45.102)$$

where ε^{IJKL} is the totally antisymmetric tensor in the internal space ($\varepsilon^{0123} = +1$) and γ is the *Barbero–Immirzi parameter*.

Why the Holst term is classically irrelevant. The Holst term $\propto \varepsilon^{IJ}{}_{KL} F_{\mu\nu}^{KL}$ is dual to the Palatini term: it replaces F^{IJ} by its internal Hodge dual $*F^{IJ} = \frac{1}{2} \varepsilon^{IJ}{}_{KL} F^{KL}$. When the torsion vanishes (the equation of motion for ω), the dual curvature $e_I^\mu e_J^\nu *F_{\mu\nu}^{IJ}$ reduces to $\varepsilon^{\mu\nu\rho\sigma} R_{\mu\nu\rho\sigma}$, which is a total derivative (the Gauss–Bonnet integrand). It does not contribute to the equations of motion. Therefore: the Holst action is classically equivalent to the Palatini action for any value of γ .

Why γ matters quantum-mechanically. The parameter γ enters the canonical analysis: it determines the relationship between the connection variable A_a^i and the geometrical variables (spin connection Γ and extrinsic curvature K). Different γ give canonically inequivalent theories: the symplectic structure (the Poisson brackets) depends on γ . At the quantum level: γ sets the scale of the area spectrum. It is analogous to the θ -parameter in QCD: a topological term that is classically irrelevant but quantum-mechanically significant.

45.24 The 3 + 1 decomposition and Ashtekar variables

To pass from the covariant Holst action to a canonical (Hamiltonian) formulation, decompose spacetime into space + time: foliate the four-dimensional spacetime $\mathcal{M}$ into spatial slices Σ_t labelled by a time parameter t .

The ADM decomposition. The metric in the 3 + 1 split: $ds^2 = -N^2 dt^2 + q_{ab}(dx^a + N^a dt)(dx^b + N^b dt)$, where N is the lapse (the rate of proper time per coordinate time), N^a is the shift (the rate of spatial coordinate change per coordinate time), and q_{ab} is the spatial 3-metric on Σ_t .

The canonical variables in the ADM formalism are (q_{ab}, π^{ab}) , where $\pi^{ab} = \frac{\sqrt{q}}{16\pi G}(K^{ab} - q^{ab}K)$ is the momentum conjugate to q_{ab} and K_{ab} is the extrinsic curvature of Σ_t .

From ADM to Ashtekar. Ashtekar’s insight (1986): reformulate the canonical variables in terms of an $SU(2)$ connection instead of the spatial metric. The starting point is the Holst action in the time gauge $e_a^0 = 0$ (the internal time direction $I = 0$ is aligned with the normal to Σ_t).

The spatial triad e_a^i ($i = 1, 2, 3$): the restriction of the tetrad to the spatial slice. It encodes the spatial metric: $q_{ab} = e_a^i e_b^j \delta_{ij}$.

The *densitised triad*:

$$E_i^a = |e| e_i^a = \sqrt{q} e_i^a, \quad (45.103)$$

where e_i^a is the inverse triad ($e_i^a e_b^i = \delta_b^a$). This is a vector density: it transforms as a vector under spatial diffeomorphisms and as a vector under internal $SU(2)$ rotations.

The *Ashtekar–Barbero connection*:

$$\boxed{A_a^i = \Gamma_a^i(e) + \gamma K_a^i}, \quad (45.104)$$

where $\Gamma_a^i(e)$ is the spin connection of the spatial slice (determined by the triad e_a^i , the analogue of the Christoffel symbols for the 3-metric) and $K_a^i = K_{ab} e^{bi}$ is the extrinsic curvature with one index converted to the internal frame.

The Barbero–Immirzi parameter γ controls the “mix” of the spin connection (a purely spatial, geometric quantity) and the extrinsic curvature (encoding the embedding of Σ_t in spacetime). At $\gamma = \pm i$: the connection A_a^i is the self-dual (or anti-self-dual) part of the 4D spin connection — Ashtekar’s original variable (1986), which simplifies the constraint algebra but has complex-valued connections. At real γ : the connection is real — Barbero’s modification (1995), which simplifies quantisation at the cost of a more complicated constraint algebra.

The Poisson brackets. The canonical pair (A_a^i, E_j^a) has the Poisson bracket:

$$\{A_a^i(\mathbf{x}), E_j^b(\mathbf{y})\} = 8\pi\gamma G \delta_j^i \delta_a^b \delta^3(\mathbf{x} - \mathbf{y}). \quad (45.105)$$

This is the same structure as in Yang–Mills theory: A_a^i is the configuration variable (the gauge connection), and E_i^a is the conjugate momentum (the “electric field”). The factor γ enters because the symplectic structure depends on the Holst term.

45.25 Quantisation: holonomies and fluxes

The passage from classical to quantum gravity follows the same logic as the quantisation of gauge theories in section ??: promote the canonical variables to operators and impose the commutation relations.

The key subtlety: the connection A_a^i is a distributional object (a 1-form), and its value at a single point is not a well-defined observable. The well-defined objects are:

Holonomies. The holonomy of A along a path e :

$$h_e[A] = \mathcal{P} \exp\left(\int_e A_a^i \tau_i dx^a\right) \in SU(2), \quad (45.106)$$

where $\tau_i = -i\sigma_i/2$ are the $SU(2)$ generators (with σ_i the Pauli matrices) and $\mathcal{P}$ denotes path ordering. The holonomy is a group element, not a Lie-algebra element: it is the $SU(2)$ analogue

of the Wilson line in lattice gauge theory (section 41.10). It encodes the integrated connection along the path — the “parallel transport” of an $SU(2)$ spinor from one end to the other.

Fluxes. The flux of E through a surface $\mathcal{S}$:

$$E_i(\mathcal{S}) = \int_{\mathcal{S}} E_i^a n_a d^2x, \quad (45.107)$$

where n_a is the normal to $\mathcal{S}$. The flux is the $SU(2)$ analogue of the electric flux through a surface: it measures the “amount of triad” (equivalently, the area element) of $\mathcal{S}$.

The holonomy-flux algebra:

$$[\hat{E}_i(\mathcal{S}), \hat{h}_e[A]] = 8\pi\gamma\ell_P^2 \tau_i \hat{h}_e[A] \cdot \kappa(e, \mathcal{S}), \quad (45.108)$$

where $\kappa(e, \mathcal{S}) = \pm 1$ or 0 depending on whether the edge e intersects the surface $\mathcal{S}$ (and on the orientation). This is the LQG analogue of $[\hat{E}, \hat{U}] = \hat{U}$ in lattice gauge theory.

45.26 Spin networks: the kinematic Hilbert space

The kinematic Hilbert space $\mathcal{H}_{\text{kin}}$ of LQG is spanned by *spin network states*: graphs embedded in the spatial manifold Σ , with quantum numbers assigned to edges and vertices.

A spin network $|\Gamma, \{j_e\}, \{i_v\}\rangle$ consists of:

A *graph* Γ : a collection of edges (1-dimensional curves) meeting at vertices (0-dimensional points), embedded in Σ .

Edge labels j_e : each edge e carries a half-integer $j_e \in \{0, \frac{1}{2}, 1, \frac{3}{2}, \dots\}$, labelling an irreducible representation of $SU(2)$ of dimension $2j_e + 1$. The representation is assigned to the holonomy $h_e[A]$ along that edge.

Vertex labels i_v : at each vertex v where n edges meet, the n representations $j_{e_1}, \dots, j_{e_n}$ must be coupled to an $SU(2)$ -invariant state (a singlet). The intertwiner i_v is a map $i_v : \bigotimes_{k=1}^n V_{j_{e_k}} \rightarrow \mathbb{C}$ that is invariant under the simultaneous $SU(2)$ action on all edges. It is the analogue of the Clebsch–Gordan coefficient for n angular momenta coupled to total $J = 0$.

The wave function. A spin network state is a functional of the connection A :

$$\Psi_{\Gamma, \{j_e\}, \{i_v\}}[A] = \bigotimes_v i_v \cdot \bigotimes_e D^{(j_e)}(h_e[A]), \quad (45.109)$$

where $D^{(j)}(g)$ is the Wigner matrix (the matrix elements of the group element g in the spin- j representation) and the intertwiners contract the matrix indices at each vertex.

The spin network states form an orthonormal basis of $\mathcal{H}_{\text{kin}}$ (the Ashtekar–Lewandowski measure on the space of connections). This is the LQG analogue of the Fock basis: a complete, labelled set of states for the quantum gravitational field.

45.27 Area and volume quantisation

The area of a surface $\mathcal{S}$, classically $A(\mathcal{S}) = \int_{\mathcal{S}} \sqrt{q_{ab} n^a n^b} d^2x$, becomes an operator $\hat{A}(\mathcal{S})$ built from the flux operators $\hat{E}_i(\mathcal{S})$.

The area operator. Using $q_{ab} = E_i^a E_j^b E_{ij}$ (where E_{ij} is the inverse of $E_i^a E_j^a$) and regularising the product of fluxes at the same point:

$$\boxed{\hat{A}(\mathcal{S}) = 8\pi\gamma\ell_P^2 \sum_{p \in \mathcal{S} \cap \Gamma} \sqrt{j_p(j_p + 1)}}, \quad (45.110)$$

where the sum runs over all points p where an edge of the spin network Γ pierces the surface $\mathcal{S}$, and j_p is the spin label of that edge.

The area spectrum is *discrete*: it takes values $A = 8\pi\gamma\ell_P^2 \sum_p \sqrt{j_p(j_p + 1)}$, a sum of $\sqrt{j(j + 1)}$ contributions. The minimum nonzero area (a single edge with $j = 1/2$):

$$A_{\min} = 8\pi\gamma\ell_P^2 \sqrt{\frac{1}{2} \cdot \frac{3}{2}} = 4\pi\gamma\ell_P^2 \sqrt{3} \approx 4\pi\gamma \times (1.6 \times 10^{-35} \text{ m})^2. \quad (45.111)$$

The volume operator. The volume of a region $\mathcal{R}$: $V(\mathcal{R}) = \int_{\mathcal{R}} \sqrt{|\det E|} d^3x$. The volume operator acts nontrivially only at the vertices of the spin network (where three or more edges meet): its eigenvalues depend on the intertwiners i_v and the spins j_e of the edges meeting at each vertex. The volume spectrum is also discrete, with a gap above zero that depends on γ .

The physical picture. A spin network state describes a discrete quantum geometry:

Each *edge* (labelled by j_e) represents a quantum of area: a surface transverse to the edge has area $8\pi\gamma\ell_P^2 \sqrt{j_e(j_e + 1)}$.

Each *vertex* (labelled by i_v) represents a quantum of volume: a region surrounding the vertex has a volume determined by i_v .

The spin network is a discrete, combinatorial structure that encodes the geometry of space. The labels (j_e, i_v) are the quantum numbers of the gravitational field, analogous to the occupation numbers $(n_{\mathbf{k}})$ of a scalar field Fock space.

45.28 Fixing the Barbero–Immirzi parameter

The Barbero–Immirzi parameter γ enters the area spectrum as an overall scale: $A \propto \gamma$. Its value is not determined by the classical theory (the Holst term is topological classically) but can be fixed by demanding consistency with the Bekenstein–Hawking entropy.

The entropy of a black hole of area A in LQG is computed by counting the number of spin-network states compatible with a horizon of area A (the “isolated horizon” boundary condition). The result:

$$S_{\text{LQG}} = \frac{\gamma_0}{\gamma} \cdot \frac{A}{4\ell_P^2}, \quad (45.112)$$

where γ_0 is a specific number determined by the combinatorics of the counting (the dominant contribution comes from edges with $j = j_{\min}$). Demanding $S_{\text{LQG}} = S_{\text{BH}} = A/(4\ell_P^2)$ fixes $\gamma = \gamma_0$. The value: $\gamma_0 = \ln 2/(\pi\sqrt{3}) \approx 0.2375$ (using the original counting with $j_{\min} = 1/2$).

This is both a success (LQG reproduces the Bekenstein–Hawking entropy with the correct $A/(4\ell_P^2)$ scaling) and a limitation (the parameter γ must be chosen by hand to match the semiclassical result, rather than being predicted).

45.29 Dynamics: constraints and spin foams

The Hamiltonian formulation of GR has no true Hamiltonian: the “Hamiltonian” is a sum of constraints that generate gauge transformations (time reparametrisations and spatial diffeomorphisms). The physical states satisfy:

$$\hat{G}_i |\Psi\rangle = 0, \quad \hat{D}_a |\Psi\rangle = 0, \quad \hat{H} |\Psi\rangle = 0, \quad (45.113)$$

where $\hat{G}_i$ is the Gauss constraint ($SU(2)$ gauge invariance), $\hat{D}_a$ is the diffeomorphism constraint (spatial diffeomorphism invariance), and $\hat{H}$ is the Hamiltonian constraint (the Wheeler–DeWitt equation).

The Gauss constraint: $\hat{G}_i |\Psi\rangle = 0$. This is $SU(2)$ gauge invariance: it requires the spin-network state to be gauge-invariant at every vertex. Spin networks with $SU(2)$ -invariant intertwiners satisfy this constraint automatically. This is the easiest constraint, already solved by the kinematic Hilbert space construction.

The diffeomorphism constraint: $\hat{D}_a |\Psi\rangle = 0$. This requires invariance under spatial diffeomorphisms: the state depends only on the equivalence class of the graph Γ up to smooth deformations. The solution: identify spin-network states that differ by spatial diffeomorphisms. The resulting “s-knot” states (equivalence classes of spin networks) span the diffeomorphism-invariant Hilbert space $\mathcal{H}_{\text{diff}}$.

The Hamiltonian constraint: $\hat{H} |\Psi\rangle = 0$. This is the hard part. The classical Hamiltonian constraint involves the curvature F_{ab}^i of the connection (a local, distributional object) acting on the spin-network states (a combinatorial, non-local structure).

Thiemann (1996) proposed a regularised Hamiltonian constraint that acts on spin-network vertices by:

- (i) Adding a new edge (a “loop”) attached to the vertex.
- (ii) Modifying the spin labels of the existing edges.
- (iii) Changing the intertwiner at the vertex.

The result: $\hat{H}$ maps one spin-network state to a superposition of other spin-network states with modified local structure around each vertex.

Whether this construction reproduces GR in the semiclassical limit — whether the coherent states peaked on a classical geometry evolve according to the Einstein equation, whether the graviton propagator emerges, whether the low-energy S -matrix matches perturbative quantum gravity — remains open.

Spin foams: the covariant approach. The spin foam programme defines the dynamics covariantly, as a sum over histories of spin networks. A spin foam is a 2-complex $\mathcal{J}$: a collection of faces (2D surfaces), edges (1D curves), and vertices (0D points), with:

Faces labelled by $SU(2)$ representations j_f .

Edges labelled by intertwiners i_e .

Vertices labelled by transition amplitudes A_v .

The partition function:

$$Z = \sum_{\mathcal{J}} \sum_{\{j_f, i_e\}} \prod_f (2j_f + 1) \prod_v A_v(j_f, i_e), \quad (45.114)$$

where the sum is over all 2-complexes and all labellings. The vertex amplitude A_v is the key dynamical ingredient: the EPRL/FK model (Engle–Pereira–Rovelli–Livine / Freidel–Krasnov) defines A_v as a specific function of the spins and intertwiners, derived from the Holst action by discretising the path integral on a simplicial complex.

The spin foam partition function is the LQG analogue of the Feynman path integral: it sums over all possible “histories” of the quantum geometry (all possible 2-complexes with all possible labellings), weighted by the vertex amplitudes. Whether this sum converges, reproduces the semiclassical limit, and defines a consistent quantum theory of gravity is the central open problem.

Part III: The Fock-Born Examination

45.30 The diagnostic question

The Fock-Born framework provides a single diagnostic for any proposed theory of quantum gravity: *which layer does the metric live in?*

If $g_{\mu\nu}$ is a Fock-layer degree of freedom (an operator on the Hilbert space, with quantum fluctuations and a propagator): the metric is part of the engine. Its eigenvalues are Fock-layer quantum numbers; its dynamics is governed by the Schrödinger equation.

If $g_{\mu\nu}$ is a Born-layer output (a macroscopic classical field, rendered by the Born rule from a non-gravitational engine): the metric is part of the rendering. It emerges from the Fock-layer correlators through the holographic dictionary, the induced-gravity mechanism, or some other bridge formula.

Both string theory and loop quantum gravity place $g_{\mu\nu}$ in the Fock layer — and both encounter characteristic difficulties that can be traced to this choice. We propose that these difficulties are not technical but structural: they arise from treating a Born-layer object as a Fock-layer input.

45.31 String theory: the graviton as a Born-layer object

String theory provides a Fock space built from the string oscillators α_n^μ , $\bar{\alpha}_n^\mu$ acting on a vacuum $|0; p\rangle$. The graviton is a specific state: $\epsilon_{\mu\nu}\alpha_{-1}^\mu\bar{\alpha}_{-1}^\nu|0; p\rangle$. The genus expansion computes scattering amplitudes as Fock-layer matrix elements. The Fock space is well-defined, the amplitudes are UV-finite, and the low-energy limit reproduces general relativity.

These are genuine achievements. But the theory has structural features that, in the Fock-Born framework, are symptoms of layer confusion.

Symptom 1: background dependence. The worldsheet sigma model is defined on a given background $(G_{\mu\nu}, B_{\mu\nu}, \Phi)$. The background metric $G_{\mu\nu}$ enters as a coupling constant of the worldsheet CFT. In the Fock-Born language: a Born-layer quantity (the classical metric) has been inserted into the Fock layer as an input. The one-way valve is violated: the rendering is recycled into the engine.

The beta-function equations $\beta_{\mu\nu}^G = 0$ (section ??) derive the Einstein equation as a self-consistency condition on the recycled input. This is elegant but circular: the engine depends on its own output. The circularity produces the landscape: many self-consistent backgrounds exist ($\sim 10^{500}$ by current estimates), and the theory cannot select among them.

Diagnosis: the background metric is a Born-layer object being used as a Fock-layer input. The multiplicity of backgrounds is the consequence of recycling: many different renderings are self-consistent, because the self-consistency equation ($\beta = 0$) has many solutions.

Symptom 2: the string is an effective description. Section ?? argued that the string is an emergent collective mode that dissolves at the lattice scale. The Polchinski–Strominger coefficient $(26 - D) = 22$ at $D = 4$ signals maximal non-fundamentality. The UV finiteness is a consequence of dissolution, not of fundamental short-distance structure.

If the string is an emergent pattern in the Born-layer rendering — a collective mode of lattice degrees of freedom, like the QCD flux tube is a collective mode of gluons — then building a Fock space from string oscillators is building a Fock space from Born-layer variables. The string Fock space $\mathcal{F}_{\text{string}}$ is a coarsened description: it captures the low-energy physics (the graviton, the gauge bosons, the matter fields) but dissolves at the lattice scale.

Diagnosis: the string oscillators are Born-layer collective modes promoted to Fock-layer degrees

of freedom. The promotion works at low energies (where the string is certified) and fails at high energies (where the string dissolves).

Symptom 3: the graviton scatters on a fixed background. String scattering amplitudes compute the S -matrix of gravitons propagating on a fixed background $(G_{\mu\nu}, B_{\mu\nu}, \Phi)$. The graviton is a small perturbation $h_{\mu\nu}$ of the background: $g_{\mu\nu} = G_{\mu\nu} + h_{\mu\nu}/M_P$. The S -matrix describes graviton-graviton scattering, graviton emission and absorption, and graviton exchange — all on the fixed stage $G_{\mu\nu}$.

In the Fock-Born framework: the background $G_{\mu\nu}$ is the Born-layer rendering (the macroscopic metric), and the graviton $h_{\mu\nu}$ is a small fluctuation of the rendering. The S -matrix computes the scattering of fluctuations in the rendering. This is the same as computing the scattering of surface waves on an ocean: the waves are real, the S -matrix is well-defined, but the ocean surface is a rendering of the molecular dynamics below, and the wave S -matrix does not describe the physics of individual water molecules.

What works: AdS/CFT. The AdS/CFT correspondence (section 45.13) resolves all three symptoms. The Fock-layer engine is the boundary CFT (non-gravitational, background-independent, well-defined). The bulk geometry is the Born-layer rendering (emergent from CFT correlators). The graviton is a Born-layer fluctuation (a mode of the rendering). The one-way valve is respected: the CFT does not use the bulk metric as an input.

AdS/CFT is string theory with the layers correctly assigned. The difficulties of perturbative string theory (background dependence, landscape, dissolution) are absent because the Fock-layer engine (the CFT) is non-gravitational. The successes (UV finiteness, graviton scattering, black hole entropy) are retained because the Born-layer rendering (the bulk) inherits them from the CFT's consistency.

45.32 Loop quantum gravity: geometry as a Born-layer object

LQG takes the opposite approach: it quantises the spatial geometry directly, with no background metric. The Hilbert space is spanned by spin networks, the area and volume are operators with discrete spectra, and the dynamics is defined by the Hamiltonian constraint $\hat{H}|\Psi\rangle = 0$.

The background independence is a genuine virtue. But the programme encounters difficulties that, in the Fock-Born framework, have a common origin: the spatial metric has been promoted from a Born-layer output to a Fock-layer operator.

Symptom 1: the problem of time. The Hamiltonian constraint $\hat{H}|\Psi\rangle = 0$ says that the physical states have zero energy: there is no time evolution in the physical Hilbert space. The “frozen” dynamics is a direct consequence of diffeomorphism invariance: if the theory is invariant under time reparametrisations, there is no preferred time, and the Hamiltonian generates a gauge transformation rather than a physical evolution.

In the Fock-Born framework: a functioning Fock-layer engine requires $H \neq 0$. The Schrödinger equation $\dot{h}_t = -iHh_t$ needs a non-trivial Hamiltonian to evolve the hidden state. The constraint $\hat{H} = 0$ produces a frozen engine: no evolution, no dynamics, no Born-layer rendering.

Diagnosis: the problem of time arises because the metric has been promoted to the Fock layer. If $g_{\mu\nu}$ is a Fock-layer operator, then diffeomorphism invariance (a gauge symmetry of $g_{\mu\nu}$) becomes a constraint on the Fock-layer state. But diffeomorphism invariance is a property of the Born-layer rendering (the metric), not of the Fock-layer engine (the lattice). On a lattice, there is a preferred time (the lattice time direction), a non-trivial Hamiltonian $H(a)$, and a functioning Schrödinger equation. The problem of time evaporates: it is an artefact of promoting a Born-layer symmetry (diffeomorphism invariance) to the Fock layer.

Symptom 2: the semiclassical limit. LQG has not yet demonstrated that its dynamics reproduces the Einstein equation in the semiclassical limit. Coherent spin-network states peaked on a classical geometry q_{ab} exist, but whether the Hamiltonian constraint evolves them according to $G_{\mu\nu} = 8\pi G T_{\mu\nu}$ is unproven. The graviton propagator, the Newtonian potential, and the linearised Einstein equation have not been derived from the full LQG dynamics.

Diagnosis: the semiclassical limit is hard because the theory is trying to recover a Born-layer equation (the Einstein equation) from a Fock-layer dynamics (the Hamiltonian constraint acting on spin networks). If the Einstein equation is a property of the rendering (induced by matter loops, as in section 43.13), then trying to derive it from a Fock-layer dynamics of the metric is trying to derive a property of the screen from the dynamics of the screen’s pixels. The screen pixels (the spin-network labels) encode the geometry, but the equation governing the geometry ($G_{\mu\nu} = 8\pi G T_{\mu\nu}$) is a rendering-level equation, not a pixel-level equation.

Symptom 3: the absence of matter. LQG is primarily a theory of pure gravity: the coupling to matter fields is technically possible but inelegant. The Standard Model fields are added by hand (fermions on the spin-network vertices, gauge fields on the edges), without a principle that determines the matter content.

Diagnosis: in the Fock-Born framework, the matter fields *are* the Fock-layer engine. The metric is induced by the matter (section 43.13). A theory that quantises the metric without the matter is quantising the rendering without the engine. The matter is not an afterthought; it is the source of everything.

What works: discrete spectra. The discrete area and volume spectra are consistent with the lattice-first viewpoint: on a lattice with spacing a , geometric observables (areas, volumes) computed from the lattice degrees of freedom naturally take discrete values proportional to powers of a . LQG’s prediction $A_{\min} \propto \gamma \ell_P^2$ is compatible with $a \sim \ell_P$. The discreteness is a Born-layer property: it is the granularity of the rendering, visible at the resolution scale of the engine.

45.33 The common diagnosis

Both approaches share the same structural error, diagnosed by the one-way valve:

	String theory	Loop quantum gravity
What is promoted	Background metric $G_{\mu\nu}$ recycled as sigma-model coupling	Spatial metric q_{ab} promoted to Fock-layer operator
Valve violation	Born output $\rightarrow$ Fock input	Born output $\rightarrow$ Fock operator
Consequence	Background dependence; landscape	Problem of time; frozen dynamics
Root cause	Metric is a rendering, not an input	Metric is a rendering, not an operator
What dissolves	The string (at a)	The spin network (at a)
What survives	AdS/CFT (CFT is the engine)	Discrete spectra (compatible with lattice)

The difficulties are different in detail (background dependence vs. frozen dynamics) but identical in origin: the metric is being treated as a Fock-layer object when it is a Born-layer output.

The graviton is a Born-layer collective mode. In the Fock-Born reading: the graviton is not a fundamental particle. It is a collective mode of the matter fields on the lattice, exactly as

the phonon is a collective mode of the atoms in a crystal. It is certified (section 44.6) by a sharp propagating pole in the stress-tensor correlator at $p^2 = 0$ (massless, spin-2). Its certification holds at energies $E \ll M_P$ (where the pole is sharp) and fails at $E \sim M_P$ (where the graviton self-energy is of order the energy, and the pole broadens).

The graviton’s certification failure at $E \sim M_P$ is the Born-layer diagnostic of the same layer confusion: the graviton is an effective particle, valid at low energies, that dissolves at the lattice scale. Trying to build a fundamental Fock space from graviton operators (string theory) or from geometry operators (LQG) is trying to build a fundamental theory from an effective degree of freedom.

The spin network is a Born-layer description. If the spatial metric is a Born-layer output, then the spin network — the basis of the Hilbert space that encodes the spatial metric — is a Born-layer description of the geometry, not a Fock-layer description of the engine. The edge labels j_e (which determine the area) and the vertex labels i_v (which determine the volume) are Born-layer quantum numbers: they describe the rendered geometry, not the engine’s internal state.

On a lattice: the engine is a collection of matter fields on sites and gauge fields on links. The “area” of a surface is a Born-layer observable computed from the lattice correlators. Its eigenvalues are discrete (because the lattice is discrete) and Planck-scale (because $a \sim \ell_P$). The spin network’s discrete area spectrum is the lattice’s rendering of spatial geometry — a Born-layer output, not a Fock-layer input.

45.34 The resolution: the engine is not gravitational

The Fock-Born framework proposes a specific resolution: the Fock-layer engine is *not gravitational*. It contains no metric, no graviton, no spin connection, no curvature. It is a lattice of matter fields (scalars, fermions, gauge fields) with a non-trivial Hamiltonian $H(a)$ that generates unitary time evolution $\dot{h}_t = -iHh_t$.

The metric $g_{\mu\nu}$ is a Born-layer output: a macroscopic classical field rendered by the Born rule from the matter correlators. The Einstein equation is induced by matter loops (the a_1 Seeley–DeWitt coefficient of the heat kernel, section 43.13). Newton’s constant is determined by the matter content and the lattice spacing: $1/G \propto N_{\text{species}}/a^2$.

The graviton is a certified collective mode: a pole in the stress-tensor correlator, with mass zero (protected by the rendered diffeomorphism invariance) and spin two (from the tensor structure of $T_{\mu\nu}$). The graviton propagator, the Einstein equation, and the graviton scattering amplitudes are all Born-layer predictions of the non-gravitational engine.

Why the difficulties dissolve.

The problem of time: the engine has a preferred time (the lattice time direction) and a non-trivial Hamiltonian $H(a)$. The Schrödinger equation $\dot{h}_t = -iHh_t$ generates real dynamics. Diffeomorphism invariance is a Born-layer symmetry of the rendered metric, not a gauge symmetry of the engine. The engine is not frozen.

Background dependence: the engine does not use $g_{\mu\nu}$ as an input. The metric is an output. There is no landscape because there is no background to choose: the engine is defined on a fixed, flat lattice.

UV finiteness: the lattice has a finite number of degrees of freedom per unit volume ($\sim 1/a^4$). No UV divergences. The graviton scattering amplitudes, computed as Born-layer correlators, are finite because the engine has no modes above π/a .

The semiclassical limit: the Einstein equation is derived, not postulated. It is the a_1 coefficient of

the heat kernel (section 43.13): a one-loop result of matter fields on the lattice. The Newtonian potential, the graviton propagator, and the graviton scattering amplitudes follow from the matter correlators by standard quantum field theory.

Matter coupling: the matter fields are not added to gravity; gravity is induced by matter. The Standard Model fields are the engine. The matter content determines G , Λ , and the higher-curvature corrections.

What is inherited. The proposed architecture inherits the achievements of both approaches while avoiding their difficulties:

Property	String	LQG	Lattice-first
UV finiteness	✓	(✓)	✓ (by construction)
Graviton	✓	?	✓ (collective mode)
Einstein equation	✓	?	✓ (induced, a_1)
Background independence	×	✓	✓ (flat lattice)
Problem of time	avoided	present	absent ($H \neq 0$)
Discrete spectra	×	✓	✓ (lattice)
Matter coupling	possible	added	built in (matter IS engine)
One-way valve	violated	violated	respected

The lattice-first column is conjectural: it describes the architecture, not a worked-out theory. Whether a specific lattice Hamiltonian $H(a)$ (with the Standard Model matter content at low energies) produces the correct graviton propagator, the correct Newton constant, and the correct cosmological constant is an open computational problem. AdS/CFT is the proof of concept: a non-gravitational engine (the CFT) whose rendering produces a gravitational bulk. The question is whether a similar architecture works for the real universe, with $\Lambda > 0$ and without the AdS boundary.

The deepest lesson. String theory and LQG are not wrong. They are Born-layer descriptions being asked to serve as Fock-layer engines. The string correctly describes the collective modes of the engine (the graviton, the gauge bosons, the matter fields) at energies below the dissolution scale. LQG correctly describes the Born-layer geometry (the discrete area and volume spectra) at the Planck scale. Both are valid as renderings. Neither is the engine.

The engine is simpler: a flat lattice, matter fields, a Hamiltonian. Everything else — the metric, the curvature, the graviton, the string, the spin network, the Einstein equation — is rendered.

46 The Lattice-First View from Computational Science

46.1 Examining QFT problems on the lattice

This subsection re-examines the traditional “problems” of quantum field theory — UV completion, non-renormalisability, the hierarchy problem, the cosmological constant problem, asymptotic safety — from the lattice-first viewpoint. All of them dissolve, partially or completely, when stated in the correct language.

The two layers, restated

The Fock layer is a computational substrate. The engine is a lattice Hamiltonian $H(a)$ acting on a Fock space $\mathcal{F}_a$, with lattice spacing a . Any computational substrate is discrete: it has a finite number of degrees of freedom per unit volume ($\sim 1/a^d$), a maximum momentum (π/a), and a maximum energy per mode ($\sim \pi/(a)$ in natural units). The engine is not an approximation to a continuum theory; it IS the theory.

The Born layer is a rendering interface. The observer accesses the engine through Born-layer measurements at a resolution $\lambda \gg a$. The rendering is the set of expectation values, correlation functions, and probability distributions computed from $H(a)$ and observed at wavelength λ . The rendering-stability condition requires only:

$$\mathcal{O}_{\text{interface}} = \mathcal{O}_{\text{physical}} + o(a/\lambda). \quad (46.1)$$

The interface outputs are insensitive to the engine’s internal structure at scales below a .

The engine has parameters $\theta(a)$ (the couplings of the lattice Hamiltonian). The interface has observables $\mathcal{O}_{\text{phys}}$ (the masses, cross sections, and thermodynamic quantities that a detector reports). The relationship between them is a function:

$$\mathcal{O}_{\text{phys}} = f(\theta(a), a), \quad (46.2)$$

computed non-perturbatively by diagonalising $H(a)$ or by evaluating the Euclidean path integral on the lattice.

Terminology

The traditional names are misleading. We translate:

Traditional name	Lattice-first name	What it is
“Bare parameter”	Engine parameter $\theta(a)$	A coupling in the lattice Hamiltonian
“Physical parameter”	Interface observable $\mathcal{O}_{\text{phys}}$	What the detector reports
“Renormalization”	Reparametrisation of the engine	Changing variables $\theta(a) \rightarrow \theta'(a)$
“Radiative correction”	A term in the perturbative expansion of f	The n -th Taylor coefficient of $f(\theta, a)$ in powers of the couplings
“Running coupling”	$\theta(a)$ as a function of a	The curve the engine parameter traces as the lattice spacing varies
“UV divergence”	A term $\propto 1/a^n$ in the Taylor expansion of f	Large Taylor coefficient; the function f itself is finite
“UV completion”	Does f have a limit as $a \rightarrow 0$?	A question about the rendering’s infinite-resolution limit
“Asymptotic safety”	Does $\theta(a)$ approach a fixed point as $a \rightarrow 0$?	A question about the engine parameter’s extrapolation

What is a “radiative correction”? The function $f(\theta, a)$ maps engine parameters to interface observables. It is computed non-perturbatively on the lattice (by Monte Carlo or exact diagonalisation), and the result is a single finite number. No “corrections” appear.

A “radiative correction” is a term in the Taylor expansion of f in powers of the coupling constants:

$$f(\theta, a) = f^{(0)}(\theta, a) + \underbrace{f^{(1)}(\theta, a)}_{\text{“one-loop correction”}} + \underbrace{f^{(2)}(\theta, a)}_{\text{“two-loop correction”}} + \dots \quad (46.3)$$

Each term $f^{(n)}$ is a “radiative correction” at n -th order. The name “radiative” is historical: in QED, the corrections come from virtual photon emission and absorption (radiation). In the lattice-first framework: they are Taylor coefficients of a smooth function, with no special physical status.

Large engine parameters are not problematic

The engine operates on a lattice with spacing a . Some engine parameters are naturally of order 1 in lattice units ($\theta \sim O(1)$ in units of a). When expressed in physical units (GeV), these become large:

$$\theta_{\text{phys}} \sim \frac{c}{a^n} \sim \left(\frac{1}{a}\right)^n \sim M_P^n \quad (\text{if } a \sim \ell_P). \quad (46.4)$$

For $n = 2$: $\theta \sim M_P^2 \sim 10^{36} \text{ GeV}^2$. For $n = 4$: $\theta \sim M_P^4 \sim 10^{72} \text{ GeV}^4$. These are large numbers in GeV, but they are $O(1)$ in lattice units. They are no more mysterious than Avogadro’s number $N_A \approx 6 \times 10^{23}$: a large number that characterises the scale of the system, not a fine-tuning or a problem.

The lattice has $\sim (1/a)^d$ sites per unit physical volume. In $d = 4$ with $a \sim \ell_P$: $\sim 10^{140}$ sites per cm^4 . Engine parameters that scale as powers of $1/a$ are simply measuring the lattice in physical units. The “largeness” is a unit conversion, not a physics problem.

Perturbative cancellations are artefacts

The “30-digit cancellation” of the hierarchy problem and the “123-digit cancellation” of the cosmological constant problem are properties of the perturbative expansion of f , not of the function f itself.

The analogy. Compute e^{-100} by its Taylor series: $\sum_{n=0}^{\infty} (-100)^n / n!$. The terms are individually of order 10^{40} ; the sum is $\sim 10^{-44}$. A “40+44 = 84 digit cancellation.” But e^{-100} is a single smooth function evaluated at a point. No cancellation is involved in its existence.

Similarly: $m_{H,\text{phys}}^2 = f(m_H^2(\text{bare}), \lambda, y_t, g, g', a)$ is a single smooth function. The perturbative expansion $f = m_H^2(\text{bare}) + c\pi^2/a^2 + O(\lambda \ln(1/a))$ presents f as a sum of large terms. But the function f itself is computed non-perturbatively on the lattice: one number in, one number out. No large terms appear. No cancellation occurs.

The perturbative expansion is a Taylor series around a badly chosen point (weak coupling, $\lambda \rightarrow 0$) that is far from the physical regime. The large Taylor coefficients ($\sim 1/a^2$, $\sim 1/a^4$) reflect the distance from the expansion point to the physical point, not a property of the function at the physical point.

The lattice computation never sees the expansion. On the lattice: the Higgs mass is extracted from the exponential decay of the correlator $\langle H(\mathbf{x})H(0) \rangle \sim e^{-m_H|\mathbf{x}|}$. The cosmological constant (if gravity is emergent) is the a_0 Seeley–DeWitt coefficient, a single computed number. No Taylor series is involved. No “radiative corrections” are computed. The function f is evaluated directly, in one step, without expansion.

The “cancellations” exist only in the perturbative language. They are self-inflicted artefacts of a computational method that is never used on the lattice.

The lattice has a maximum energy

The lattice has a maximum momentum $|\mathbf{k}| \leq \pi/a$ (the edge of the Brillouin zone) and a maximum energy per mode $\omega_{\max} \sim \pi/a$. No physical process on the lattice can produce a centre-of-mass energy $\sqrt{s} > \pi\sqrt{d}/a$ (the energy of a mode at the corner of the Brillouin zone in d dimensions).

Taking $\sqrt{s} \rightarrow \infty$ is unphysical. The Mandelstam variable s (section 19.4) is bounded by the lattice:

$$s \leq s_{\max} \sim \frac{\pi^2 d}{a^2}. \quad (46.5)$$

Scattering amplitudes $\mathcal{A}(s, t)$ are defined only for $s \leq s_{\max}$. The high-energy limit “ $s \rightarrow \infty$ ” is a property of the continuum extrapolation, not of the lattice.

This has immediate consequences:

The Froissart bound ($\sigma_{\text{tot}} \leq C \ln^2(s/s_0)$) is automatically satisfied on the lattice: s is bounded, so σ_{tot} is bounded.

The Regge limit ($s \rightarrow \infty$ at fixed t) is an extrapolation beyond the lattice’s domain. The Regge trajectories (section 41.9) are valid at $s \ll s_{\max}$ and terminate at the lattice scale.

String theory’s UV finiteness (the Gross–Mende exponential softening at $s \rightarrow \infty$, section 45.7) describes the behaviour of the rendered amplitude beyond the lattice scale. On the lattice: there are no scattering events at $s > s_{\max}$. The exponential softening is a property of the analytic continuation of the effective description, not of the engine (section ??).

The only genuine concern: numerical stability

Every “problem” listed above is either a perturbative artefact, a question about terminology, or a question about the mathematical properties of the rendering’s infinite-resolution limit. None is a physical concern about the engine.

There is one genuine physical concern: *numerical stability* — whether a small change in an engine parameter $\theta(a)$ produces a large change in an interface observable $\mathcal{O}_{\text{phys}}$.

The condition number. Define the sensitivity:

$$\kappa = \frac{\partial \ln \mathcal{O}_{\text{phys}}}{\partial \ln \theta} = \frac{\theta}{\mathcal{O}} \frac{\partial \mathcal{O}}{\partial \theta}. \quad (46.6)$$

If $|\kappa| \sim O(1)$: a 1% change in θ produces a $\sim 1\%$ change in $\mathcal{O}$. Stable.

If $|\kappa| \gg 1$: a 1% change in θ produces a $\gg 1\%$ change in $\mathcal{O}$. Numerically unstable. The engine must be calibrated with high precision.

The origin of large κ : phase transitions. The function $f(\theta, a)$ maps engine parameters to interface observables. Near a phase transition: f has a singularity (a critical point), and κ diverges.

For the Higgs sector: the lattice ϕ^4 theory in $d = 4$ has a critical surface separating the symmetric phase ($\langle \phi \rangle = 0$) from the broken phase ($\langle \phi \rangle \neq 0$). The physical Higgs mass is:

$$m_{H, \text{phys}} \sim \frac{1}{a} |m_H^2(\text{bare}) - m_{H,c}^2|^{1/2}, \quad (46.7)$$

where $m_{H,c}^2$ is the critical bare mass. The condition number:

$$\kappa = \frac{m_H^2(\text{bare})}{m_{H,\text{phys}}^2} \frac{\partial m_{H,\text{phys}}^2}{\partial m_H^2(\text{bare})} \sim \frac{m_H^2(\text{bare})}{m_H^2(\text{bare}) - m_{H,c}^2}. \quad (46.8)$$

Near the critical surface ($m_H^2(\text{bare}) \approx m_{H,c}^2$): $\kappa \rightarrow \infty$. The engine is numerically unstable: a tiny change in the bare mass produces a large change in the physical mass.

This is the non-perturbative content of the hierarchy problem: not a cancellation between large perturbative terms, but a large condition number near a critical surface.

When is a symmetry relevant? A symmetry “protects” a mass by making $\kappa = O(1)$ (stable) or $\kappa = 0$ (exactly stable):

Gauge symmetry (photon): the critical surface for $m_\gamma \neq 0$ does not exist within the gauge-invariant theory. The condition number is $\kappa = 0$: the photon mass is exactly zero regardless of the engine parameters. No numerical instability.

Chiral symmetry (fermion at $m = 0$): the critical surface is at $m_{\text{bare}} = 0$ by symmetry. Near $m_{\text{bare}} = 0$: $m_{\text{phys}} \propto m_{\text{bare}}$, so $\kappa = 1$. Stable.

No symmetry (Higgs): the critical surface is at $m_{H,c}^2 \neq 0$, and the distance to it is a free parameter. Near the critical surface: $\kappa \gg 1$. Numerically unstable.

The question “why is the Higgs light?” becomes “why is the engine operating near the critical surface of the Higgs sector?” — a question about the engine’s phase-diagram location, not about cancellations in a Taylor series.

Examination of the traditional “problems”

We now examine each traditional problem and classify it as: **(A)** a perturbative artefact, **(B)** a question about the rendering’s mathematical properties, or **(C)** a genuine concern about numerical stability.

1. UV divergences in QFT. *Traditional statement:* loop integrals diverge as $\int d^d k k^n \rightarrow \infty$.

Lattice-first diagnosis: **A.** On the lattice, all integrals are sums over finitely many modes ($|k_i| \leq \pi/a$). Every quantity is finite. The “divergences” are properties of the continuum extrapolation $a \rightarrow 0$, not of the engine at any finite a .

2. Non-renormalisability of gravity. *Traditional statement:* the perturbative expansion of quantum gravity produces new divergences at each loop order, requiring infinitely many counterterms.

Lattice-first diagnosis: **A.** On the lattice, there are no divergences and no counterterms. The “new divergences at each loop order” are new terms in the Taylor expansion of f that grow as higher powers of $1/a$. The function f itself is finite. If gravity is emergent (section 43.13), the gravitational couplings are not engine parameters at all but interface outputs (a_0, a_1 coefficients), and the question is moot.

3. The hierarchy problem. *Traditional statement:* the Higgs mass receives “quadratically divergent radiative corrections” $\delta m_H^2 \sim 1/a^2$, requiring a 30-digit cancellation.

Lattice-first diagnosis: **A** (the cancellation) + **C** (the numerical instability).

The “30-digit cancellation” is a perturbative artefact (section 41.11): it exists only in the Taylor expansion of f , not in the function itself. On the lattice, the Higgs mass is computed non-perturbatively from a single correlator. No large terms, no cancellation.

The non-perturbative content is the numerical instability (46.8): the engine operates near the critical surface of the Higgs sector ($m_H \cdot a \ll 1$ implies $|m_H^2(\text{bare}) - m_{H,c}^2| \ll 1/a^2$), and the condition number $\kappa \gg 1$. This is a genuine structural property of the engine, independent of perturbation theory.

4. The cosmological constant problem. *Traditional statement:* the QFT vacuum energy $\rho_\Lambda^{\text{QFT}} \sim 1/a^4$ disagrees with the observed $\rho_\Lambda^{\text{obs}} \sim (2.3 \text{ meV})^4$ by 123 orders of magnitude.

Lattice-first diagnosis: **A** (the “ 10^{123} ”) + possibly **dissolved** (if gravity is emergent).

The “ 10^{123} ” ratio compares a perturbatively computed vacuum energy ($\sum \frac{1}{2}\omega_k$ summed to $k = \infty$ in the continuum) to a measured number. On the lattice, the vacuum energy is a single finite number (the ground-state energy of $H(a)$), and the “ 10^{123} mismatch” does not arise.

If gravity is emergent: Λ is not an engine parameter. It is the a_0 Seeley–DeWitt coefficient — an interface output computed from the matter content. There is no bare Λ_0 in the engine, no cancellation between Λ_0 and $\delta\Lambda$, and no fine-tuning. The cosmological constant is whatever the matter vacuum energy renders.

If gravity is NOT emergent (if Λ_0 is an engine parameter): then $\kappa \sim 10^{120}$ (the most extreme numerical instability in physics), and the concern is **C**.

5. Asymptotic safety. *Traditional statement:* does the gravitational coupling $\tilde{g}(a) = G(a)/a^2$ approach a non-trivial fixed point as $a \rightarrow 0$?

Lattice-first diagnosis: **B**. Asymptotic safety is a question about whether the engine parameter curve $\theta(a)$ has a fixed point in the $a \rightarrow 0$ limit. If yes: the continuum limit exists, and the theory has finitely many free parameters. If no: no continuum limit, but the engine at finite a is unchanged.

The physical content: if the fixed point exists, it determines the economy of description (how many inputs the engine needs). It does not affect the interface outputs at $\lambda \gg a$.

6. UV completion. *Traditional statement:* does the theory have a well-defined high-energy limit? Can it be extended to arbitrarily short distances?

Lattice-first diagnosis: **B**. UV completion asks whether the rendering can be refined to infinite resolution ($a \rightarrow 0$ with the physics unchanged). The engine operates at finite a and does not need a $a \rightarrow 0$ limit. UV completion is a question about the mathematical properties of the rendering’s extrapolation, not about the physics at any finite a .

7. String theory’s UV finiteness. *Traditional statement:* string scattering amplitudes are UV-finite at every loop order, thanks to modular invariance.

Lattice-first diagnosis: **B** (partially). On the lattice, all amplitudes are finite because $s \leq s_{\text{max}} \sim \pi^2 d/a^2$. The string’s UV finiteness is a property of the rendered effective description at $s \ll s_{\text{max}}$: the modular invariance of the worldsheet CFT ensures that the analytic continuation of the amplitude beyond s_{max} is finite. But on the lattice, $s > s_{\text{max}}$ does not exist; the “UV region” is not part of the engine.

If the string is an emergent collective mode that dissolves at the lattice scale (section ??): its UV finiteness is caused by the dissolution, not by a fundamental property of short-distance physics.

8. The Landau pole in QED. *Traditional statement:* the QED coupling $\alpha(\mu)$ diverges at $\mu = \Lambda_L \sim e^{3\pi/(2\alpha)} \sim 10^{286}$ GeV (the Landau pole).

Lattice-first diagnosis: **B.** On the lattice, $\alpha(a)$ is a finite engine parameter at every a . The Landau pole is a feature of the perturbative beta function extrapolated to $\mu \rightarrow \infty$ (equivalently, $a \rightarrow 0$): the Taylor expansion of $\alpha(\mu)$ diverges at $\mu = \Lambda_L$. Whether the non-perturbative function $\alpha(a)$ actually diverges at some finite a is a lattice question (triviality of QED): the evidence suggests it does not diverge but instead approaches zero (the theory is “trivial” — free in the continuum limit). Either way: the physics at $\mu \ll \Lambda_L$ is unaffected.

9. The strong CP problem. *Traditional statement:* why is $\theta_{\text{QCD}} \approx 0$?

Lattice-first diagnosis: **C** (partially). θ_{QCD} is an engine parameter (a coefficient in the lattice Hamiltonian), and the experimental bound $\theta < 10^{-10}$ requires it to be very close to zero. This is a numerical fine-tuning in the engine parameters, but θ is an angular variable (range $[0, 2\pi)$), so its smallness does not involve large numbers or powers of $1/a$. The condition number κ is $O(1)$ (no phase transition is involved): a small change in θ produces a small change in the neutron electric dipole moment. The “problem” is that $\theta \approx 0$ has no known symmetry explanation within the SM (unless the axion, which provides the symmetry, exists).

The summary table

	“Problem”	Type	Lattice-first status
1	UV divergences	A	Finite on the lattice; artefact of $a \rightarrow 0$
2	Non-renormalisability	A	No counterterms on the lattice; G, Λ are outputs if gravity is emergent
3	Hierarchy problem	A+C	Cancellation is perturbative artefact; numerical instability near critical surface is genuine
4	Cosmological constant	A (or dis-solved)	“ 10^{123} ” is perturbative; dissolved if gravity is emergent (Λ is not a bare param.)
5	Asymptotic safety	B	Economy of description; does not affect physics at finite a
6	UV completion	B	Rendering’s infinite-resolution limit; engine runs at finite a
7	String UV finiteness	B	Property of the rendered description; engine is finite by $s \leq s_{\text{max}}$
8	Landau pole	B	Perturbative extrapolation; non-perturbative $\alpha(a)$ is finite
9	Strong CP	C	$\theta \approx 0$ is unexplained but $\kappa \sim O(1)$; not a large-number problem

Type A: perturbative artefact (does not exist on the lattice). Type B: question about the rendering (does not affect the engine). Type C: genuine numerical stability concern (exists on the lattice, independent of perturbation theory).

The only entries with type C content are:

The hierarchy problem: the engine operates near a critical surface ($\kappa \gg 1$), which is a non-perturbative property of the phase diagram.

The cosmological constant (only if gravity is not emergent): $\kappa \sim 10^{120}$, the most extreme

numerical instability.

The strong CP problem: $\theta \approx 0$ is unexplained, but $\kappa \sim O(1)$ (no phase transition involved), so it is a parameter fine-tuning, not a numerical instability.

Everything else is either a perturbative artefact or a question about the rendering’s mathematical properties. The engine is finite, well-defined, and produces stable interface outputs. The “problems” are problems of the continuum extrapolation, not of the physics.

46.2 Quantum foundation and HEP viewed from computational sciences

The language of this book — engine, rendering, interface, lattice, Fock layer, Born layer — is not standard in physics. But it is the native language of a large and rapidly growing community: the practitioners of computational statistics, machine learning, and artificial intelligence. Their perspective on the “deep problems” of fundamental physics is instructive, and occasionally devastating.

What they already know

The computational sciences have been borrowing from statistical mechanics for decades:

Markov chain Monte Carlo (Metropolis et al. 1953): sample from a probability distribution $P(x) \propto e^{-E(x)/T}$ by constructing a Markov chain whose stationary distribution is P . This is the Boltzmann distribution of a statistical-mechanical system at temperature T , and the algorithm is identical to lattice Monte Carlo (section 23.5).

Variational inference (Jordan et al. 1999): approximate a complex distribution $P(x)$ by the closest member of a tractable family $Q_\theta(x)$, minimising the KL divergence $D_{\text{KL}}(Q\|P)$. This is the variational principle of quantum mechanics — the trial wave function $|\psi_\theta\rangle$ that minimises $\langle H \rangle_\theta$ — transplanted to probability distributions.

Diffusion models (Sohl-Dickstein et al. 2015, Ho et al. 2020): learn to reverse a diffusion process that gradually destroys data by adding noise. The forward process is a stochastic differential equation (Langevin dynamics); the reverse process is the score function $\nabla_x \ln P(x)$. The mathematics is identical to stochastic quantisation (Parisi and Wu, 1981): the Euclidean path integral is the stationary distribution of a Langevin process in a fictitious time.

Normalising flows (Rezende and Mohamed, 2015): transform a simple distribution (Gaussian) into a complex one through a sequence of invertible maps. This is the renormalisation group flow (Wilson, 1974): a sequence of coarse-graining transformations that maps the microscopic Boltzmann distribution to the macroscopic effective distribution.

Tensor networks (White 1992, Vidal 2007): represent a high-dimensional wave function as a network of low-rank tensors. This is the density-matrix renormalisation group (DMRG), developed for condensed-matter systems and now used in machine learning for dimensionality reduction.

These are not analogies. They are the same mathematics, applied to different systems. A machine-learning researcher who understands MCMC, variational methods, and diffusion models already understands the computational core of quantum field theory. The engine-interface distinction — a computational process (the engine) generating data (the interface) that the user observes — is second nature to them.

What would puzzle them about quantum foundations

A researcher from machine learning, encountering the measurement problem of quantum mechanics for the first time, might react with bewilderment — not at the physics, but at the

framing.

The measurement problem asks: why does the deterministic Schrödinger equation (the engine) produce probabilistic outcomes (the interface)? How does the wave function “collapse”? What selects a specific outcome from the superposition?

From the computational perspective: every simulation has an engine and an interface. The engine runs deterministically (the game loop, the physics solver, the neural network forward pass). The interface renders the output for the user (the screen, the audio, the text token). The user sees discrete, definite outputs (a pixel colour, a word, a measurement result). The engine maintains a continuous, deterministic internal state (the hidden variables, the activations, the wave function).

Nobody asks “why does the game engine produce definite pixel colours?” or “how does the neural network collapse its activations into a single output token?” The answer is obvious: the interface is a *projection* of the internal state onto the user’s observables. The projection discards information (the engine state has more degrees of freedom than the interface output). The discarded information appears as randomness to the user.

The Born rule $P(\lambda) = \|P_\lambda h\|^2$ is exactly this projection: the engine state h is projected onto the eigensubspace P_λ of the measured observable, and the squared norm gives the probability. The “collapse” is the update of the user’s knowledge upon receiving the interface output. The engine state does not collapse; it continues to evolve unitarily. The “measurement problem” is the question of why the interface exists — which, from the computational perspective, is like asking why the monitor exists. It is not a scandal; it is the design.

The decades of debate about interpretations of quantum mechanics — Copenhagen, many-worlds, Bohmian mechanics, consistent histories, QBism — look, from this perspective, like debates about the relationship between the game engine and the monitor. They are interesting philosophical discussions, but they do not affect the engine’s code, the interface’s behaviour, or the user’s experience.

What would puzzle them about high-energy physics

A computational scientist, reading the high-energy physics literature, would encounter three practices that, from their perspective, are puzzling:

1. Worrying about sub-lattice physics. The lattice has spacing a . The interface renders at resolution $\lambda \gg a$. The rendering-stability condition guarantees that the interface outputs are insensitive to the engine’s structure below a . Yet a large fraction of the HEP literature is devoted to the question: “what happens at distances shorter than a ?”

UV completion, asymptotic safety, string theory at the Planck scale, trans-Planckian physics, the information paradox at the singularity — all of these ask about physics below the engine’s resolution. From the computational perspective: this is like asking what happens between the pixels of a GPU, or what the game engine computes between clock cycles. The answer is: nothing. The engine has a finite resolution. Below that resolution, the question is not well-posed.

The lattice-first viewpoint makes this explicit: the engine operates at spacing a , and a is the smallest meaningful length. “What happens at $r < a$?” is not a physics question; it is a question about the mathematical properties of the continuum extrapolation (the rendering’s infinite-resolution limit). The engine does not have sub-lattice structure, just as a digital simulation does not have sub-pixel physics.

2. Worrying about large Taylor coefficients. The function $f(\theta, a)$ maps engine parameters to interface observables. It is a smooth, finite, well-defined function, computed non-perturbatively

on the lattice. Yet the HEP literature spends enormous effort on the *Taylor expansion* of f and is alarmed when the Taylor coefficients are large ($\sim 1/a^2$ for the Higgs, $\sim 1/a^4$ for the cosmological constant).

A computational scientist would recognise this as a *numerical analysis* issue, not a physics issue. If you Taylor-expand a smooth function around a badly chosen point, the coefficients can be arbitrarily large. The function e^{-100} has Taylor coefficients of order 10^{40} . The function $\sin(10^6 x)$ has Taylor coefficients of order 10^{6n} at order n . In no case does this indicate a problem with the function — only with the expansion.

The “hierarchy problem” and the “cosmological constant problem” are, from this perspective, complaints about the numerical properties of a particular Taylor expansion of a perfectly well-behaved function. The solution is not to find new physics (SUSY, extra dimensions, the multiverse) but to stop Taylor-expanding and compute the function directly. The lattice does this.

3. Building the engine from screen objects. This is the deepest puzzle. Both string theory and loop quantum gravity attempt to construct the Fock-layer engine from objects that belong to the Born-layer interface.

String theory builds the engine from gravitons — fluctuations of the spacetime metric $g_{\mu\nu}$, which is an interface observable (what a ruler and a clock report). The worldsheet sigma model uses the metric as an input ($G_{\mu\nu}$ appears as a coupling constant of the 2D CFT). The metric is a rendered quantity — what the user sees on the screen — being recycled as an engine component.

Loop quantum gravity builds the engine from spin networks — quantum states of the spatial geometry, which is also an interface observable (what a surveyor reports). The area and volume operators measure the rendered geometry and promote it to a Fock-layer operator.

From the computational perspective: this is like building a game engine out of the rendered graphics. The 3D world that the player sees on the monitor is not the engine; it is the engine’s output. You cannot build the engine from the pixels on the screen. You cannot run the physics solver using the rendered image as input. The engine and the rendering are different computational objects, with different data structures, different algorithms, and a one-way data flow (engine $\rightarrow$ rendering, not rendering $\rightarrow$ engine).

The AdS/CFT correspondence (section 45.13) is the one example in high-energy physics where the engine-interface distinction is correctly implemented: the boundary CFT is the engine (non-gravitational, well-defined), and the bulk geometry is the rendering (emergent, gravitational). The bulk metric is not an input to the CFT; it is computed from the CFT’s correlators. The one-way valve is respected.

The computational scientist would ask: why isn’t this the default architecture for quantum gravity? Why do string theory and LQG insist on building the engine from screen objects, when AdS/CFT has demonstrated that the correct architecture is a non-gravitational engine with a gravitational rendering?

A note on Dirac

Of all the physicists of the twentieth century, it is Dirac for whom the author feels the deepest admiration — and the deepest sympathy.

The admiration is for the work. Dirac’s leap from the Poisson bracket $\{A, B\}_{\text{PB}}$ to the commutator $[\hat{A}, \hat{B}]/(i\hbar)$ was not a derivation but an act of physical intuition so precise that it created quantum mechanics in its mature form. The *transformation theory* (1927) unified Heisenberg’s matrix mechanics and Schrödinger’s wave mechanics into a single framework — the Hilbert space

formulation that remains the foundation of the Fock layer to this day. The quantisation of the electromagnetic field (1927) created quantum field theory. The Dirac equation (1928) — derived not from experiment but from the requirement that quantum mechanics be consistent with special relativity — predicted antimatter before it was observed, gave the electron its magnetic moment to leading order, and remains, in the author’s view, the most beautiful equation in physics. The hole theory, reinterpreted as the Dirac sea and then as antiparticles, was the first glimpse of the Fock space structure: the vacuum is not empty but full, and particles and antiparticles are excitations above and below the Fermi level.

Dirac did all of this before the age of 30.

The sympathy is for the struggle that consumed his later decades. Dirac never accepted renormalization. He called it “sweeping the infinities under the rug” and spent the last thirty years of his life searching for a formulation of QED that would be finite without subtractions. He tried new Hamiltonians, new variational principles, new representations of the Lorentz group. He never found one.

From the lattice-first viewpoint: Dirac’s diagnosis was correct but his search was misdirected. The infinities are not a defect of the dynamics (the Hamiltonian, the Lagrangian, the equations of motion). They are a defect of the *continuum* ($\mathbb{R}^{3,1}$ as the arena of physics). On a lattice with spacing a : the self-energy of the electron is $E_{\text{self}} = \sum_{|k| < \pi/a} k^2 / (k^2 + m^2) < \infty$, a finite sum. The Lamb shift is a finite difference between two finite numbers. The anomalous magnetic moment is a finite ratio. Every quantity that Schwinger, Tomonaga, and Feynman computed by subtracting infinity from infinity is, on the lattice, a straightforward finite computation.

The renormalization programme that Dirac rejected is, in the lattice-first reading, the demonstration that the computational substrate is discrete. The “infinities” are what happen when you pretend the lattice spacing is zero ($a \rightarrow 0$) and the momentum integration is unbounded ($\int d^4k$ to infinity). The “subtraction” is what you do to undo this pretence and recover the finite lattice answer. Renormalization is not a trick; it is the correction of an error (the continuum assumption) by an *ad hoc* procedure (the subtraction) that happens to give the right answer (the lattice answer).

Dirac was right that the procedure is unsatisfying. But the resolution was not a new dynamics — it was a new *substrate*. The engine is discrete. It has parameters $\theta(a)$, a lattice spacing a , and a rendering-stability condition $\mathcal{O}_{\text{phys}} = f(\theta(a), a) + o(a/\lambda)$. We can allow a to vary (the renormalization group), but asking what happens at $a = 0$ (the continuum limit) is asking the engine to operate without a substrate. The engine can be refined ($a \rightarrow a/2$, then $a \rightarrow a/4$, and so on), and the rendering improves, but the limit $a \rightarrow 0$ is a mathematical extrapolation, not a physical requirement.

Had Dirac known Wilson’s lattice (1974), he might have found peace: the infinities that tormented him for thirty years are artefacts of a continuum that does not exist. The engine is finite. It always was.

The view, summarised

From the perspective of the computational sciences:

The engine-interface distinction is obvious. Every simulation, every neural network, every MCMC sampler has an internal state (the engine) and an output (the interface). The user sees the output, not the engine. The output is a projection of the internal state, with information loss. The information loss appears as randomness. This is the Born rule.

Sub-lattice questions are ill-posed. The engine has a finite resolution (a). Asking “what happens below a ?” is asking about the engine’s internal clock cycles between output frames. The answer

is: the engine computes, but the computation has no “spatial” or “temporal” interpretation below the lattice scale. UV completion is the question of whether the engine can be refined to infinite resolution. It is an interesting mathematical question, but the engine runs fine at finite resolution.

Large Taylor coefficients are a numerical analysis issue. Don’t expand. Compute. The lattice computes f directly. The Taylor expansion is a convenience for analytic understanding, not a necessity for computation. If the expansion has large coefficients, choose a better expansion (or don’t expand at all).

Build the engine from engine objects, not screen objects. The rendered graphics (the spacetime metric, the graviton, the geometry) are screen objects. The engine is something else — a lattice, a matrix model, a non-gravitational quantum system. The screen objects emerge from the engine through the rendering pipeline. They are not the engine’s building blocks.

These are not profound insights. They are the common sense of the computational sciences, applied to physics. The lattice-first viewpoint of this book is, in essence, the computational common sense that the engine comes first and the rendering comes second.

46.3 The two Clay problems: a lattice-first reading

Two of the seven Clay Millennium Prize Problems concern physics: the *Navier–Stokes existence and smoothness* problem and the *Yang–Mills existence and mass gap* problem. Both ask whether a continuum theory (a PDE on $\mathbb{R}^3$ or $\mathbb{R}^4$) has certain mathematical properties (smooth solutions, or a spectral gap). From the lattice-first viewpoint: both ask the wrong question. The physics is in the discrete substrate, not in the continuum rendering.

The parallel is illuminated by a historical episode that predates both problems by a century.

Einstein and Brownian motion

In 1905, Einstein studied the random motion of a pollen grain suspended in water. His goal was not to describe the motion but to *prove that molecules exist*: to show that the observed jitter of the grain is caused by the bombardment of discrete water molecules, and to predict a measurable quantity (the diffusion coefficient $D = k_B T / (6\pi\eta r)$) that Perrin could test.

The physics has two scales:

The engine: water molecules, with a mean free path $a \sim 10^{-8}$ cm and a collision time $\tau_c \sim 10^{-12}$ s. The molecular dynamics is a deterministic Hamiltonian system of $\sim 10^{23}$ particles.

The rendering: the pollen grain, observed at resolution $\lambda \sim 10^{-4}$ cm and time intervals $\Delta t \sim 10^{-2}$ s. The grain’s trajectory looks random because the observer cannot resolve the individual molecular collisions.

Einstein’s result: at the rendering scale ($\Delta t \gg \tau_c$, $\lambda \gg a$), the grain’s displacement satisfies $\langle x^2 \rangle = 2D\Delta t$, and the diffusion coefficient D is related to the molecular parameters (temperature, viscosity, grain radius) by the Einstein relation. Perrin’s measurements (1908) confirmed the prediction and established the reality of molecules. Nobel Prize 1926.

The essential point: the continuum description (diffusion equation, Brownian paths) is the *rendering*. The discrete description (molecular collisions) is the *engine*. Einstein’s achievement was to infer the engine from the rendering.

Wiener’s measure: rigorous but wrong physics

In 1923, Wiener constructed a rigorous probability measure on the space of continuous paths $x(t) \in C([0, T], \mathbb{R}^d)$: the *Wiener measure*. It is the mathematical foundation of stochastic analysis, the basis of the Itô calculus, and the starting point for the rigorous theory of diffusion processes.

The Wiener measure describes Brownian motion as a continuous path that is nowhere differentiable: the velocity dx/dt does not exist at any time t . This is a beautiful mathematical object. But from Einstein’s perspective, it is the *wrong physics*.

Einstein’s whole point was that the motion is NOT continuous at the molecular scale. The pollen grain receives discrete kicks from individual molecules, separated by the collision time $\tau_c \sim 10^{-12}$ s. Between kicks: the grain moves ballistically (smooth, differentiable trajectory). The “continuous, nowhere-differentiable” character of the Wiener path is an artefact of the rendering at $\Delta t \gg \tau_c$: when the individual kicks are unresolved, the trajectory *looks* continuous and rough. But it isn’t. The engine is discrete.

The Wiener measure makes the rendering rigorous. But in doing so, it discards the engine. The continuous path $x(t)$ is a mathematical idealisation of the molecular trajectory — an idealisation that Einstein would have found ironic, since his entire argument was designed to prove that the molecular (discrete, non-continuous) description is the real one.

	Engine (discrete)	Rendering (continuum)
Brownian motion	Molecular collisions ($a \sim 10^{-8}$ cm)	Wiener process ($x(t) \in C[0, T]$)
Physics proved by	Einstein (1905): molecules exist	Wiener (1923): measure on paths exists
Resolution	$\Delta t \sim \tau_c \sim 10^{-12}$ s	$\Delta t \gg \tau_c$: rendering scale
Differentiable?	Yes (ballistic between kicks)	No (nowhere differentiable)
Status	The physics	The mathematical idealisation

Navier–Stokes: the first Clay problem

The Navier–Stokes equations describe the velocity field $\mathbf{u}(\mathbf{x}, t)$ of an incompressible viscous fluid on $\mathbb{R}^3$:

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\nabla p + \nu \nabla^2 \mathbf{u}, \quad \nabla \cdot \mathbf{u} = 0. \quad (46.9)$$

The Clay problem asks: given smooth initial data $\mathbf{u}_0$, does a smooth solution exist for all time $t > 0$? Or can singularities (blow-up of $|\nabla \mathbf{u}|$ in finite time) develop?

From the lattice-first viewpoint: the Navier–Stokes equation is the *rendering* of molecular dynamics. It is obtained by coarse-graining the molecular Hamiltonian (the Boltzmann equation $\rightarrow$ the Chapman–Enskog expansion $\rightarrow$ the Navier–Stokes equation), discarding the molecular structure at scales below the mean free path a .

The question “does the Navier–Stokes equation have smooth solutions for all time?” is a question about the rendering: does the continuum description remain self-consistent at all times, or does it break down (develop a singularity)?

If the rendering breaks down (a Navier–Stokes singularity forms): this does NOT mean the physics breaks down. The molecular dynamics (the engine) continues to evolve smoothly — molecules do not stop obeying Newton’s laws because the continuum approximation fails. A Navier–Stokes singularity is a failure of the rendering, not of the engine. It signals that the continuum description has been extrapolated beyond its domain of validity: the velocity gradient

$|\nabla \mathbf{u}|$ has become so large that the variation occurs on scales comparable to the mean free path, where the continuum assumption fails.

The Clay problem, in the lattice-first reading, asks: is the rendering self-consistent at all resolutions? Can the continuum description be refined to infinite resolution without encountering a singularity? This is a question about the mathematical properties of the Navier–Stokes PDE, not about the physics of fluids. The physics is in the molecules. The molecules are fine.

Yang–Mills mass gap: the second Clay problem

The Clay problem asks (Jaffe and Witten, 2000):

- (i) Prove that a quantum Yang–Mills theory exists on $\mathbb{R}^4$ satisfying the Wightman axioms (a Hilbert space, a unitary Poincaré representation, fields as operator-valued distributions, locality, and the spectral condition).
- (ii) Prove that it has a mass gap $\Delta > 0$.

From the lattice-first viewpoint: this problem has exactly the same structure as the Navier–Stokes problem and the Wiener measure.

	Engine (discrete)	Rendering (continuum)
Brownian motion	Molecular collisions	Wiener measure on $C[0, T]$
Fluid dynamics	Molecular dynamics (Hamiltonian)	Navier–Stokes on $\mathbb{R}^3$
Yang–Mills theory	Lattice gauge theory (Wilson, $H(a)$)	Wightman QFT on $\mathbb{R}^4$
Clay problem asks	—	Does the continuum rendering exist and have specific properties?
Physics lives in	The engine	—

The Wightman axioms are the Yang–Mills analogue of the Wiener measure: they make the continuum rendering rigorous. Constructing a Wightman QFT from the lattice theory is the Yang–Mills analogue of constructing the Wiener measure from molecular dynamics: it shows that the rendering has a mathematically well-defined infinite-resolution limit.

But just as the Wiener measure discards the molecular engine (replacing discrete kicks with continuous-nowhere-differentiable paths), the Wightman axioms discard the lattice engine (replacing link variables and plaquettes with operator-valued distributions on $\mathbb{R}^4$). And just as Einstein would have found the Wiener measure ironic (it makes rigorous the very continuum that Einstein proved is not fundamental), Wilson might find the Wightman-axiom formulation of Yang–Mills ironic: it makes rigorous the very continuum that Wilson proved is not needed.

The physically correct formulation. As argued in section 41.12: the physically meaningful question is not whether a Wightman QFT exists, but whether the lattice Yang–Mills theory has a mass gap with rendering stability:

Prove that lattice $SU(N)$ Yang–Mills theory in $d = 4$ has a mass gap $\Delta(a, L) > 0$ that survives the thermodynamic limit $L \rightarrow \infty$ and satisfies $\Delta(a) = \Delta_{\text{phys}} + O(a^2)$ as $a/\lambda \rightarrow 0$.

This formulation:

- (i) Does not require the continuum limit to exist.
- (ii) Does not require the Wightman axioms (which encode the continuum assumption).
- (iii) Does not require exact Poincaré invariance (a rendered symmetry, valid at $\lambda \gg a$).

(iv) Does require the physically meaningful content: a spectral gap on the lattice, uniform in volume, stable under refinement.

The lattice formulation is to the Clay formulation as Einstein's molecular theory is to the Wiener measure: it is the engine-level statement of the same physics, without the continuum idealisation.

The pattern

Before stating the pattern, a remark on the mathematical tools.

The Lebesgue–Wiener analogy. The Lebesgue integral extends the Riemann integral to handle pathological functions (the Dirichlet function, the characteristic function of the rationals). It greatly simplifies the mathematical infrastructure: the dominated convergence theorem, Fubini's theorem, and the L^p spaces are all cleaner with Lebesgue. But for every function a physicist encounters, the Riemann integral gives the same answer. The Lebesgue theory is a mathematical convenience, not a physical necessity.

The Wiener measure plays the same role for stochastic processes. It provides a rigorous foundation for the Itô calculus, martingale theory, and the theory of stochastic differential equations. But for the physics of Brownian motion, it is never needed.

The Fokker–Planck equation — the continuum description of the probability density $P(x, t)$ of a diffusing particle — can be derived entirely on the lattice, without the Wiener measure:

- (i) Start with a discrete random walk on a lattice with spacing Δx and time step Δt .
- (ii) Write the master equation: $P(x, t + \Delta t) = \sum_{x'} W(x|x') P(x', t)$.
- (iii) Taylor-expand to second order in Δx : $P(x, t + \Delta t) \approx P + \langle \Delta x \rangle \partial_x P + \frac{1}{2} \langle (\Delta x)^2 \rangle \partial_x^2 P$.
- (iv) Take the expectation on the lattice (the average over the transition kernel W). No measure theory required: the expectation is a finite sum.
- (v) Take the limit $\Delta t \rightarrow 0$ (with $\langle (\Delta x)^2 \rangle / (2\Delta t) \rightarrow D$): obtain the Fokker–Planck equation $\partial_t P = -\partial_x(\mu P) + D \partial_x^2 P$.

Every step is a finite computation on a lattice. The Wiener measure enters only if one insists on formulating the process as a measure on continuous paths — which is a mathematical elegance, not a physical requirement.

Black–Scholes without Wiener. The Black–Scholes option pricing formula can be derived from the Cox–Ross–Rubinstein binomial tree: a discrete-time, discrete-state model of stock price evolution. The binomial tree converges to the Black–Scholes formula as the number of steps $N \rightarrow \infty$. The derivation uses only elementary probability (the binomial distribution) and the no-arbitrage condition. No Itô calculus, no stochastic differential equations, no Girsanov theorem, no Wiener measure.

The mathematical finance textbooks that develop the full machinery of continuous-time stochastic analysis are beautiful mathematics. But for a practitioner who wants to price an option, the binomial tree suffices: it is simpler, more intuitive, numerically stable, and produces the same answer.

The lesson generalises: the continuum formulation (Wiener measure, Itô calculus, Wightman axioms, Navier–Stokes PDE) is a mathematical convenience that simplifies general theorems and enables elegant proofs. But the physics — the actual computation of numbers that can be compared to experiment — lives on the lattice. The lattice computation is finite, explicit, and does not require the continuum limit to exist.

What if the answers are negative? The Clay problems ask whether the continuum rendering

is self-consistent. There are two possible answers:

Positive answer (the continuum rendering is self-consistent): the Navier–Stokes equation has smooth solutions for all time; the Wightman Yang–Mills theory exists with a mass gap. This would be a mathematical achievement: the construction of rigorous continuum objects from the discrete engine.

Negative answer (the continuum rendering breaks down): the Navier–Stokes equation develops singularities; the Wightman axioms cannot be satisfied for Yang–Mills in $d = 4$. This would be a *physical* achievement: it would confirm that the continuum description is an approximation with a finite domain of validity, and that the correct description is the discrete engine.

From the lattice-first viewpoint: a negative answer would be the more interesting result. It would validate Einstein’s thesis (the continuum is not fundamental; molecules are) and Wilson’s thesis (the lattice is the definition of QCD; the continuum is the rendering). A Navier–Stokes singularity would be the fluid-dynamical analogue of a lattice artefact: the continuum rendering, pushed beyond its domain of validity, produces a mathematical pathology that has no counterpart in the molecular engine. A failure of the Wightman axioms for Yang–Mills would be the field-theoretic analogue: the continuum rendering, pushed to infinite resolution, is not self-consistent — but the lattice theory is perfectly well-defined at every a .

The negative answers would not be disasters. They would be confirmations that the engine is discrete and the rendering is an approximation. The physics — the mass gap, the hadron spectrum, the confinement, the fluid flow — is computed on the lattice and is unaffected by whether the continuum limit exists. The Clay problems, in this reading, are tests of the continuum hypothesis, not of the physics.

The pattern, stated.

In all three cases — Brownian motion, fluid dynamics, Yang–Mills theory — the same pattern occurs:

Step 1: the physics is discovered in the discrete engine (Einstein’s molecules, Wilson’s lattice, molecular dynamics).

Step 2: a continuum rendering is constructed (Wiener’s measure, the Wightman axioms, the Navier–Stokes equation) that captures the long-distance physics while discarding the engine.

Step 3: mathematicians ask whether the continuum rendering is self-consistent (does the Wiener measure exist? does the Navier–Stokes equation have smooth solutions? does the Wightman QFT exist with a mass gap?).

Step 4: these questions become famous open problems, attracting enormous mathematical effort — even though the physics was settled at Step 1.

The Clay problems are Step 3 questions. They ask about the mathematical properties of the rendering, not about the physics of the engine. They are important mathematics — the construction of rigorous continuum limits is a deep and beautiful enterprise. But they are not physics questions. The physics was answered by Einstein (molecules exist, Brownian motion is their rendering), by Wilson (the lattice defines QCD, the continuum is its rendering), and by the molecular dynamicists (Newton’s laws define the fluid, the Navier–Stokes equation is its rendering).

The engine is discrete. The rendering is continuous. The Clay problems ask whether the rendering is self-consistent. The engine does not care.

46.4 The source code and the KISS principle

A computational scientist, having understood the engine-interface architecture, would ask the ultimate question: *what is the source code?*

She would follow the KISS principle (“Keep It Simple, Sir”) — the engineering maxim that the best design is the simplest one that works. Unnecessary complexity is not elegance; it is a liability. Every additional data structure, every additional symmetry group, every additional field species is a line of code that must be maintained, debugged, and justified.

Four lines

The source code of quantum physics on a finite lattice consists of four lines:

Line 1: the engine.

$$h_{t+dt} = e^{-iH dt} h_t. \quad (46.10)$$

A rotation of a vector h_t in a finite-dimensional Hilbert space $\mathcal{F}$, generated by a Hermitian operator H . The dimension of $\mathcal{F}$ is unthinkable large (exponential in the number of lattice sites: $\dim \mathcal{F} \sim d^{N_{\text{sites}}}$, where d is the local Hilbert space dimension), but finite. The rotation is unitary, deterministic, and reversible. This is the entire engine: one matrix exponential, one vector, one time step.

Lines 2–4: the rendering API.

(a) *Born probability*:

$$P(\lambda) = \|P_\lambda h_t\|^2. \quad (46.11)$$

The probability of outcome λ is the squared norm of the projection of h_t onto the eigensubspace of the measured observable $\hat{O}$ at eigenvalue λ .

(b) *Sampling*:

$$\lambda^* \sim P(\lambda). \quad (46.12)$$

Draw one sample λ^* from the distribution P . This is what the detector reports.

(c) *Trimming* (the handshake):

$$h_t \longrightarrow \frac{P_{\lambda^*} h_t}{\|P_{\lambda^*} h_t\|}. \quad (46.13)$$

After the sample is rendered, update h_t by projecting onto the measured eigensubspace and normalising. This is the “collapse” — not a physical process but a handshake between the engine and the interface: the engine’s state is trimmed to be consistent with the rendered output.

That is the entire programme. Four lines. Everything else — the Standard Model, general relativity, thermodynamics, chemistry, biology — is the *content* of H (line 1) and the *choice* of $\hat{O}$ (line 2): what the engine computes and what the interface renders.

The source code does not mention spacetime, fields, particles, symmetries, dimensions, or coupling constants. These are all properties of the specific Hamiltonian H and the specific observables $\hat{O}$ — the data, not the programme.

The KISS principle applied to H

Given that the programme is four lines, the entire complexity of physics resides in the choice of H . The KISS principle demands:

Minimal particle content. The simplest local Hilbert space is a scalar: $d = 1$ internal degree of freedom per site (a real number, or a finite set of values). A scalar field on a lattice has

H built from nearest- neighbour couplings: kinetic term (hopping between sites), mass term (on-site energy), and self-interaction (on-site potential).

KISS prefers: start with scalars. If vectors (spin-1), spinors (spin-1/2), and tensors (spin-2) can be shown to emerge as certified collective modes of the scalar substrate — as the phonon, the magnon, and the graviton emerge from atomic systems — then they need not be postulated. They are rendered, not coded.

Whether this is physically possible (whether the Standard Model’s fermions and gauge bosons can emerge from a purely scalar lattice) is an open and difficult question. But KISS says: try the simplest option first.

Minimal symmetry. Symmetries constrain H : gauge invariance forbids certain terms, Lorentz invariance dictates the form of the kinetic energy, SUSY pairs bosons with fermions. Each symmetry reduces the number of free parameters but increases the structural complexity of H .

KISS says: symmetry is a *means*, not an *end*. A symmetry earns its place only if it simplifies the engine — if the constrained H has fewer parameters and produces the observed interface outputs. A symmetry that complicates the engine (by doubling the particle content, introducing new dimensions, or requiring elaborate constructions) must be justified by a corresponding simplification of the output.

Nature itself endorses this attitude. Parity (P) is violated by the weak force: the W boson couples to left-handed fermions but not right-handed ones. Charge conjugation (C) is violated. CP is violated (by the CKM phase). Time reversal (T) is violated (by CPT and CP). Nature does not maximise symmetry; it uses exactly as much as it needs and no more.

Even the human body encodes this lesson: we have one heart, on the left side. Bilateral symmetry is approximate, not exact. The symmetry serves a function (locomotion, sensory processing), and where it serves no function (the placement of internal organs), it is broken.

Minimal interaction. The simplest interaction is a monomial of the field at a single site: ϕ^4 (the simplest non-trivial self-interaction for a real scalar), or a nearest-neighbour coupling $\phi_i\phi_j$ (the simplest interaction between sites).

KISS prefers: the lowest-order, shortest-range interaction that produces the observed physics. Higher-order terms (ϕ^6 , ϕ^8 , next-nearest-neighbour couplings) are irrelevant operators (section ??): they are suppressed at long wavelengths by powers of a/λ and do not affect the rendering. The engine can include them, but it need not.

The SUSY critique

Supersymmetry doubles the particle content: every boson gets a fermionic partner, every fermion gets a bosonic partner. The Minimal Supersymmetric Standard Model (MSSM) has 32 additional particles (squarks, sleptons, gauginos, Higgsinos), 105 new parameters (the soft SUSY-breaking terms), and a new symmetry algebra (the $\mathcal{N} = 1$ super-Poincaré algebra with its central charges).

The motivation: SUSY cancels the $1/a^2$ terms in the perturbative expansion of the Higgs mass correction. But as argued in section 41.11: these terms are artefacts of the perturbative expansion, not of the physics. On the lattice: the Higgs mass is computed non-perturbatively from a single correlator. No large terms appear. No cancellation is needed. No SUSY is required.

KISS would ask: you want to add 32 particles, 105 parameters, and a new algebraic structure to cure a large Taylor coefficient of a smooth function? The function is finite. The Taylor coefficient is large because the expansion point is far from the physical regime. The solution is to stop expanding, not to redesign the engine.

The LHC searched for superpartners at energies up to ~ 2 TeV. None were found. KISS is unsurprised.

The string critique

Consider what the superstring requires to produce a four-dimensional world with fermions:

The bosonic string: 26 dimensions, a tachyon (vacuum instability), no fermions.

Adding worldsheet supersymmetry: introduces fermionic partners ψ^μ on the worldsheet, reduces the critical dimension from 26 to 10. But the spectrum still contains a tachyon in the Neveu–Schwarz sector.

The GSO projection: projects out the tachyon and half the spectrum, selecting states with definite worldsheet fermion parity. The projection is a discrete choice (two possible projections give Type IIA and Type IIB).

Compactification: the six extra dimensions must be curled up on a Calabi–Yau manifold (a Ricci-flat Kähler threefold with $SU(3)$ holonomy) to preserve $\mathcal{N} = 1$ SUSY in 4D. The choice of CY determines the gauge group, matter content, and coupling constants. There are $\sim 10^5$ known CY manifolds.

Fluxes: quantised p -form field strengths through the cycles of the CY, needed to stabilise the moduli (the massless scalars that would otherwise mediate fifth forces). Each cycle can carry an independent flux integer.

Branes: D-branes wrapped on cycles of the CY, providing gauge symmetry ($U(N)$) and chiral matter (strings stretching between brane stacks). The brane configuration must satisfy tadpole cancellation conditions.

SUSY breaking: the vacuum must break SUSY (since superpartners are not observed), which requires an anti-brane, a non-perturbative effect, or some other mechanism.

The total: 10 dimensions, worldsheet supersymmetry, GSO projection, a Calabi–Yau manifold, quantised fluxes, D-brane stacks, and a SUSY-breaking mechanism — to produce the four-dimensional Standard Model with its three generations of fermions.

Now ask: how would you write this as source code?

```
// The string engine (pseudocode)
dim = 10;
susy = N1_worldsheet;
gso = project(sector, (-1)^F == +1);
cy = load_calabi_yau("quintic_quotient_Z5");
fluxes = quantize(cy.three_cycles, tadpole_bound);
branes = wrap(D3, cy.four_cycles, gauge_group=SU(3)xSU(2)xU(1));
susy_break = add_anti_D3(warped_throat);
landscape_vacuum = select(10^500_vacua, anthropic=true);
// ... then compute scattering amplitudes
```

A computational scientist would look at this and ask: is this really the source code of the universe? Or is this the rendering pipeline — the long chain of effective descriptions, each valid in a limited regime, patched together to produce the observed output?

KISS says: the source code should be four lines ((46.10)–(46.13)), with H a nearest-neighbour Hamiltonian on a lattice. The 10 dimensions, the Calabi–Yau, the fluxes, the branes, the GSO projection, and the landscape are all rendering-level constructs: properties of the effective string description at $\lambda \gg a$, not of the engine at the lattice scale.

The string is not the source code. It is the screen saver.

The simulation test. The critique sharpens when one asks: how would you actually *simulate* string theory on a computer?

The perturbative definition (the genus expansion) computes scattering amplitudes as integrals over moduli spaces of Riemann surfaces. These integrals can be evaluated analytically at low genus (tree level, one loop) for specific external states. But the genus expansion is asymptotic: it diverges, and no finite number of terms gives the exact answer. The non-perturbative effects ($\sim e^{-1/g_s}$, from D-brane instantons) are invisible in the expansion. You cannot simulate the full theory by summing the genus series.

The non-perturbative definition (M-theory, BFSS, AdS/CFT) does not have a computable action for general backgrounds. The BFSS matrix model *can* be simulated: it is a quantum mechanics of $N \times N$ matrices, and Monte Carlo simulations of the Euclidean version exist (Anagnostopoulos et al., Catterall and Wiseman, Berkowitz et al.). But what does the BFSS simulation compute? It evolves $N \times N$ matrices on a *lattice in time*: a one-dimensional grid with spacing Δt . The matrices are the degrees of freedom at each time step. The strings, the branes, and the geometry are *emergent patterns* in the matrix data — they are not coded into the simulation but read off from the output.

The BFSS simulation is, in essence, a lattice computation with matrices as the site variables. It is not a string simulation; it is a matrix simulation from which strings emerge. The practitioner who runs the code never manipulates strings, never sums over Riemann surfaces, never compactifies on a Calabi–Yau. They diagonalise matrices and measure eigenvalue distributions. The string is a pattern in the rendered data.

This is the simulation test: if you cannot simulate the theory without going back to a lattice (or a matrix model, or some other discrete substrate) and making the strings composite, then the strings are not the source code. They are the output. The source code is whatever runs on the lattice.

Topology change as interpretation. In perturbative string theory: topology change is fundamental. The genus expansion sums over worldsheets of different topology (sphere, torus, double torus, ...). The splitting and joining of strings — one string becoming two, two strings merging into one — is a change of worldsheet topology.

On the lattice: there is no topology change. The lattice has a fixed topology (a hypercubic grid, or whatever the substrate is). The fields evolve on this fixed grid. The “strings” are patterns in the field correlators — extended structures that look like one-dimensional objects at $\lambda \gg a$. When two such patterns merge (two flux tubes joining), or one pattern splits (a flux tube breaking), the lattice computation continues smoothly: the field values at the grid points change, but the grid itself does not change topology.

The “topology change” is an interpretation of the computational result, not the computation itself. Just as the “curved spacetime” of numerical relativity is an interpretation of the data on a flat grid (section 46.5), the “genus expansion” of string theory is an interpretation of the patterns that emerge from a fixed-topology lattice computation. The computation is on the lattice. The topology is on the screen.

The graviton and gravity. In string theory: the graviton is the spin-2 state $\epsilon_{\mu\nu}\alpha_{-1}^{\mu}\bar{\alpha}_{-1}^{\nu}|0;p\rangle$, and gravitational interactions arise from the exchange of this state. The Einstein equation is derived from the beta-function condition $\beta_{\mu\nu}^G = R_{\mu\nu} + \dots = 0$ (section ??).

But in the induced gravity framework (section 43.13): Newton’s constant receives contributions from *all* matter species through the a_1 Seeley–DeWitt coefficient: $1/(16\pi G_{\text{ind}}) =$

$N_{\text{species}}/(288\pi a^2)$. The graviton of the string is one spin-2 mode among many. The gravitational effect at macroscopic distances is dominated by the induced action (the sum over all species), not by the exchange of a single string graviton.

The string graviton may be a valid collective mode (a certified spin-2 excitation of the flux tube), but it is not necessarily the *dominant* source of gravity. The dominant source is the collective effect of all matter fields on the lattice, encoded in the heat kernel coefficient a_1 . The string graviton is a Born-layer pattern; the induced G is a Born-layer output of the full engine. The pattern may contribute to the output, but the output is not reducible to the pattern.

Ice cubes and water. An analogy clarifies the relationship. Strings are like ice cubes: rigid, structured, frozen patterns of a more fundamental substrate. Scalars (or whatever the engine's degrees of freedom are) are like water: fluid, formless, providing the medium from which the rigid patterns crystallise.

If you tried to build an ocean out of ice cubes, you would fail: the cubes are rigid and cannot flow, they freeze the dynamics, and the construction is fragile (raise the temperature and the cubes melt). But you can build ice cubes out of water: cool the water below the freezing point and the rigid structures form spontaneously. The ice is a collective mode of the water. It inherits the water's properties (density, latent heat, crystal structure) but it is not the water.

In this analogy: the string's spin-2 mode (the graviton vertex operator $\epsilon_{\mu\nu}\alpha_{-1}^\mu\bar{\alpha}_{-1}^\nu|0;p\rangle$) is the gravitational contribution of one ice cube. But the ocean's gravity comes mostly from the water — the induced Newton's constant G_{ind} from the a_1 coefficient, summed over all N_{species} matter fields, most of which are not strings but the fluid substrate from which strings form.

The scalar substrate provides a *fail-safe mode* for the string. At low energies ($E \ll 1/\ell_s$): the string is a well-defined collective mode, certified by a sharp pole in the stress-tensor correlator. Its spin-2 mode contributes to gravity. At high energies ($E \sim 1/a$): the string dissolves (section ??), the ice melts back into water, and the gravitational effect continues uninterrupted — carried by the scalar substrate through the induced a_1 coefficient, which does not require the string to exist.

The string does not fail gracefully; it simply melts. The water was always there. Gravity was always induced. The string graviton was a contribution, not the cause.

What the source code might look like

The KISS engine, speculatively:

```
// The lattice engine (pseudocode)
lattice = hypercubic(d=4, spacing=a, sites=N);
fields = scalar(lattice, components=k);
H = hopping(fields, nearest_neighbour)
  + mass(fields, m_squared)
  + interaction(fields, phi_4, coupling=lambda);
state = initial_state(fock_space(fields));
evolve(state, H, dt);
```

Five lines. A four-dimensional lattice with k -component scalar fields, nearest-neighbour hopping, a mass, and a ϕ^4 self-interaction. The gauge symmetry, the fermions, the graviton, and the spacetime geometry are conjectured to emerge from this substrate as certified collective modes — just as phonons emerge from atomic vibrations, magnons from spin interactions, and the speed of sound from the equation of state.

Whether this is too simple (whether scalar fields on a 4D lattice can produce the Standard Model)

is an open question. Perhaps the engine needs gauge fields on the links (Wilson’s construction), or fermions on the sites (the lattice SM), or both. But KISS demands that the starting point be the simplest option, and that complexity be added only when the simple option fails.

The four-line programme ((46.10)–(46.13)) is not speculative. It is the exact source code of quantum mechanics. The only question is: what is H ?

That is the only question.

46.5 Numerical relativity is more than an approximation

Numerical relativists solve the Einstein equations on a computer. They discretise spacetime on a grid (finite differences, spectral methods, adaptive mesh refinement), evolve the metric components $g_{\mu\nu}$ forward in time, and extract gravitational waveforms, black hole mergers, and neutron star collisions. The field has produced some of the most important results in modern gravitational physics: the first binary black hole waveform templates (Pretorius 2005), the prediction of the gravitational wave signal that LIGO detected in 2015, and the modelling of neutron star mergers that produce kilonovae.

The practitioners view their work as an *approximation* to general relativity. The grid is a computational convenience; the “real” theory is the continuum Einstein equation on $\mathbb{R}^{3,1}$; the grid spacing Δx is a source of error to be minimised by refinement.

From the lattice-first viewpoint: this self-description is backwards. The numerical relativists are not approximating general relativity. They are *implementing* it — on a flat, internal computational lattice, where $g_{\mu\nu}$ is a pattern generated by the lattice data. They are, unknowingly, running a classical version of the game engine.

What numerical relativity actually does

Consider a standard 3+1 numerical relativity code (e.g., the BSSN formulation):

The computational substrate. A three-dimensional Cartesian grid with spacing Δx , stored in computer memory. The grid is *flat*: the memory addresses are arranged in a regular array. The grid points are labelled by integer indices (i, j, k) , with coordinates $x^a = i \cdot \Delta x$, $y^a = j \cdot \Delta x$, $z^a = k \cdot \Delta x$. The grid has no curvature, no dynamical geometry, no metric: it is a featureless rectangular array of memory locations.

The data on the grid. At each grid point (i, j, k) : a set of floating-point numbers representing the spatial metric γ_{ab} , the extrinsic curvature K_{ab} , the lapse α , the shift β^a , and any matter fields. These are *numbers stored at grid points* — not geometric objects “living on a curved manifold.”

The evolution. The BSSN evolution equations update the data at each grid point from one time step to the next: $\gamma_{ab}(t + \Delta t) = \gamma_{ab}(t) + \Delta t \cdot \dot{\gamma}_{ab}(t)$, where $\dot{\gamma}_{ab}$ is computed from the data at neighbouring grid points (finite differences for the spatial derivatives).

The output. Gravitational waveforms are extracted at large radii by computing the Weyl scalar Ψ_4 (a specific combination of the data at grid points) and decomposing it into spherical harmonics. The waveform $h_+(t)$, $h_\times(t)$ is what LIGO measures: an interface observable.

The architecture.

	Numerical relativity	Lattice-first quantum gravity
Substrate	Flat Cartesian grid in computer memory	Flat lattice with spacing a
Data at sites	Floating-point numbers $(\gamma_{ab}, K_{ab}, \alpha, \beta^a)$	Quantum fields (ϕ, ψ, A_μ)
Evolution	BSSN equations (classical, deterministic)	$h_{t+dt} = e^{-iHdt}h_t$ (quantum, unitary)
Curvature	A pattern in the stored data	A pattern in the field correlators
Output	Gravitational waveforms $h_+, h_\times$	Born-layer observables
Status of $g_{\mu\nu}$	Numbers on a flat grid	Rendered from matter correlators

The architectures are identical. The only difference is the nature of the data at the grid points: floating-point numbers (classical) versus quantum field operators (quantum). The substrate is flat in both. The curvature is a pattern in the data, not a property of the substrate. The metric $g_{\mu\nu}$ is not the arena on which the computation takes place; it is the *result* of the computation.

Numerical relativists already know this, operationally. They never curve their computer’s memory. They never implement non-Euclidean addressing of memory locations. They store data on a flat grid and let the data encode the curvature. The “curved spacetime” is an interpretation of the numbers on the grid, not a property of the grid itself.

From floating-point numbers to quantum fields

The passage from numerical relativity to quantum gravity, in the lattice-first framework, is remarkably direct:

Step 1: replace the floating-point numbers with quantum fields. At each lattice site: instead of a classical number $\gamma_{ab}(i, j, k) \in \mathbb{R}$, place a quantum field operator $\hat{\phi}(i, j, k)$ acting on a local Hilbert space. The simplest choice: a scalar field $\hat{\phi}$ with $[\hat{\phi}(\mathbf{x}), \hat{\pi}(\mathbf{y})] = i\delta_{\mathbf{xy}}$.

Step 2: let the interactions generate the metric. The lattice Hamiltonian H contains nearest-neighbour couplings (hopping terms, self-interactions). Under the Wilson renormalisation group (section ??): integrating out short-distance fluctuations generates an effective action for the long-wavelength modes. The Seeley–DeWitt expansion (section 43.13) shows that the one-loop effective action on a curved background includes:

$$\Gamma_{\text{eff}} \supset \frac{1}{16\pi G_{\text{ind}}} \int \sqrt{g} R \quad (\text{the induced Einstein–Hilbert action}). \quad (46.14)$$

The metric $g_{\mu\nu}$ and the curvature R emerge from the scalar field correlators, not from the substrate.

Step 3: the HS connection. The Hubbard–Stratonovich transformation (section ??) introduces an auxiliary field $\sigma(x)$ for the scalar composite. In the gravitational context: the metric perturbation $h_{\mu\nu}$ is the auxiliary field for the stress-tensor composite $T_{\mu\nu}$ of the scalar fields:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \quad h_{\mu\nu} \longleftrightarrow \sigma \sim G \cdot T_{\mu\nu}[\phi]. \quad (46.15)$$

The metric perturbation $h_{\mu\nu}$ is a collective mode of the scalar field, promoted from a Born-layer observable (the stress-tensor correlator) to an effective Fock-layer field via HS. The full metric $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ is the flat substrate plus the collective pattern.

Step 4: the scalar field propagates “on g .” The effective action for ϕ in the background of its own collective mode $h_{\mu\nu}$:

$$S_{\text{eff}}[\phi, h] \supset \frac{1}{2} \int \sqrt{g} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi. \quad (46.16)$$

The scalar field appears to propagate on the curved metric $g_{\mu\nu}$, even though the underlying substrate is flat. The “curvature” is a pattern in the correlators of ϕ : the effective propagation speed, the effective light cones, and the effective geodesics are all determined by the collective mode $h_{\mu\nu}$, which is itself a function of ϕ .

The self-consistency is exact: ϕ generates $h_{\mu\nu}$ (through the HS transformation), and $h_{\mu\nu}$ modifies the propagation of ϕ (through the effective action). This circular structure IS the Einstein equation: $G_{\mu\nu} = 8\pi G T_{\mu\nu}[\phi]$, where both sides are functionals of the same underlying field ϕ on a flat lattice.

Step 5: clocks and rulers are composites. A “clock” is a composite of the scalar field: a localised oscillation mode with a definite frequency $\omega = E_1 - E_0$ (the gap between two energy levels of a bound state of ϕ). A “ruler” is a composite: the equilibrium separation between two bound states. Both are Born-layer observables, computed from the correlators of ϕ .

These composites also propagate “on $g_{\mu\nu}$ ”: their effective dynamics is governed by the same collective mode $h_{\mu\nu}$ that governs the propagation of ϕ itself. When a clock is transported around a closed path on the lattice: its accumulated phase depends on $h_{\mu\nu}$ along the path. This phase difference IS the gravitational redshift, the Shapiro delay, the geodetic precession — all the Born-layer predictions of general relativity, computed from the scalar field correlators on a flat lattice.

The metric $g_{\mu\nu}$ is the *interpretation* of the measurement data (what the composite clocks and rulers report), not the substrate on which the computation runs. Numerical relativists already do this: they interpret the data on their flat grid as a curved spacetime. The lattice-first framework says: this interpretation is not an approximation. It is the correct architecture.

Quantum gravity is simpler than numerical relativity

Numerical relativity solves the Einstein equations: 10 coupled nonlinear PDEs (the BSSN system), plus 4 gauge conditions (the lapse and shift), plus constraint equations (the Hamiltonian and momentum constraints), plus boundary conditions (excision for black holes, outgoing waves at infinity), plus adaptive mesh refinement, plus spectral filtering, plus apparent horizon tracking. A typical production code is $\sim 10^5$ – 10^6 lines.

The lattice-first quantum engine: $h_{t+dt} = e^{-iHdt} h_t$, where H is a nearest-neighbour scalar Hamiltonian on a 4D lattice. Four lines ((46.10)–(46.13)).

The Einstein equations, the BSSN system, the gauge conditions, the constraints, the boundary conditions — all of these emerge in the Born layer as properties of the effective metric $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}[\phi]$. They are not coded into the engine; they are rendered by the engine. The engine is simpler than its rendering.

This is not paradoxical. It is the universal pattern of emergence: the rules of chess (32 pieces, a few movement rules) are simpler than the space of possible games ($\sim 10^{120}$ positions). The rules of Go (one type of stone, one capture rule) are simpler than AlphaGo’s evaluation function. The lattice Hamiltonian H is simpler than the Einstein equations that emerge from it, because H contains only the engine’s local rules, while the Einstein equations encode the global, self-consistent rendering of those rules at macroscopic scales.

Numerical relativists solve the rendering. The lattice-first framework solves the engine. The engine is simpler.

47 Appendix: But What Is the Observer?

Throughout this book, the observer has appeared in every equation and been defined in none. The Born rule says: the observer submits an observable $\hat{O}$ and receives an eigenvalue λ with probability $P(\lambda) = \|P_\lambda h_t\|^2$. The Bohr collapse says: the hidden state is trimmed to the projected superposition and renormalized. But who submits? Who receives? What *is* the observer?

The Copenhagen axioms are deliberately silent. Bohr insisted—with a discipline that bordered on stubbornness—that the observer is not a concept to be defined within quantum mechanics but a precondition for its application. The classical apparatus, the outcome, the record: these are given, not derived. The axioms do not say what the observer is made of, why the observer exists, or how the observer chooses which observable to submit. They say only that if an observer exists and submits an observable, the Born rule gives the probabilities.

This silence is not evasion. It is the recognition that quantum mechanics, as a physical theory, is a theory of the relationship between the hidden state and the observed outcome—not a theory of the observer itself. The observer is the user of the theory, not its subject.

The three founders of quantum mechanics—Bohr, Heisenberg, and Schrödinger—held three different views on what lies behind this silence. Their positions illuminate the question even if they do not answer it.

47.1 Schrödinger’s journey: from realism to Vedanta

The early Schrödinger: the wave function is physical

When Schrödinger published the wave equation in 1926, his interpretation was straightforwardly realistic: the wave function $\psi(\mathbf{x}, t)$ is a physical field, as real as the electromagnetic field. The electron is not a particle; it is a wave packet—a localised excitation of ψ that propagates through space. The squared modulus $|\psi|^2$ is the charge density, not a probability. There is no collapse, no measurement problem, no observer: the wave function evolves deterministically under the Schrödinger equation, and that is all.

This interpretation was elegant, intuitive, and wrong. Lorentz and others pointed out that wave packets spread: a localised ψ disperses over time. Born showed that $|\psi|^2$ must be a probability density, not a charge density, to account for discrete scattering outcomes. And the extension to many particles required ψ to live in configuration space $\mathbb{R}^{3N}$, not physical space $\mathbb{R}^3$ —making the “physical field” interpretation untenable for $N > 1$.

Schrödinger resisted. He found the probabilistic interpretation distasteful and the collapse postulate absurd. In 1935, he invented the cat paradox as a *reductio ad absurdum*: if the wave function is a probability and the collapse is real, then a cat is in a superposition of alive and dead until someone opens the box. The paradox was intended to ridicule the Copenhagen interpretation, not to pose a genuine puzzle.

The argument with Bohr: Copenhagen, September 1926

The confrontation between Schrödinger and Bohr is one of the most dramatic episodes in the history of physics. Heisenberg, who was present, left a vivid account.

Schrödinger came to Copenhagen to defend the wave interpretation: ψ is physical, particles are wave packets, quantum mechanics is a theory of continuous waves, and the quantum jumps that Bohr had postulated since 1913 are an artefact of an incomplete description. Bohr met him at the train station and began arguing immediately.

The argument continued without pause for days—at the institute, at Bohr’s home, over meals,

into the small hours. Bohr's style was relentless: he would seize on every concession, pursue every implication, and never let a point rest until it was resolved or the interlocutor was exhausted. Schrödinger was both.

Bohr's position was unyielding: the wave function is not a physical field. The outcomes are discrete. The classical apparatus is a precondition for the quantum description, not derivable from it. The quantum jumps are real—not as events in the wave function but as transitions in the observable outcomes.

Schrödinger, worn down by the relentless argument, fell ill and took to bed in Bohr's guest room. Bohr's wife Margrethe brought him soup and tea and tended to him. Bohr sat at his bedside and continued arguing.

Heisenberg recalled Schrödinger finally exclaiming: "If one has to go on with these damned quantum jumps, then I am sorry that I ever got involved with quantum theory." To which Bohr replied: "But the rest of us are very grateful that you did, because your wave mechanics, in its mathematical clarity and simplicity, represents a gigantic progress over the previous form of quantum mechanics."

Schrödinger left Copenhagen unconvinced but shaken. He could not refute Bohr's arguments: the wave interpretation could not account for the discreteness of outcomes or the correlations of entanglement. But he could not accept the Copenhagen alternative: a theory in which the wave function is not physical and the observer is undefined.

The argument was never resolved between them. But it transformed Schrödinger. Over the next two decades, he abandoned wave realism entirely—and replaced it with something that neither Bohr nor Heisenberg could have anticipated.

The turn to Vedanta

Schrödinger's engagement with Indian philosophy was not a late-life curiosity. It began in the early 1920s, before the wave equation, through his reading of Schopenhauer (who was deeply influenced by the Upanishads) and through direct engagement with translations of the Vedantic texts. But it deepened decisively after the confrontation with Bohr, as Schrödinger sought a philosophical framework that could accommodate the quantum formalism without the Copenhagen silence or the wave-realist naivety.

He found it in *Advaita Vedanta*—the non-dualist tradition of Shankara (8th century CE), rooted in the Upanishads, the oldest philosophical texts of the Indian tradition.

The core doctrine, as Schrödinger understood it:

Consciousness is not plural. The multiplicity of minds is apparent, not real. The individual self (Atman) is identical with the universal principle (Brahman): tat tvam asi—"thou art that." The manifold world of separate objects and separate observers is maya—an appearance projected by the one consciousness upon itself.

Schrödinger did not treat this as metaphor. He adopted it as his considered philosophical position, stated with full seriousness in his published works.

In *What Is Life?* (1944)—the book that inspired a generation of molecular biologists—he devoted the epilogue to consciousness:

"The overall number of minds is just one. I venture to call it indestructible since it has a peculiar time-table, namely mind is always now. There is really no before and after for mind. There is only a now that includes memories and expectations."

In *Mind and Matter* (1958):

“The world is given to me only once, not one existing and one perceived. Subject and object are only one. The barrier between them cannot be said to have broken down as a result of recent experience in the physical sciences, for this barrier does not exist.”

In *My View of the World* (1961), his most explicitly philosophical work:

“Consciousness is a singular of which the plural is unknown. It is not something that can be divided or multiplied. There is only one thing, and what seems to be a plurality is merely a series of different aspects of this one thing, produced by a deception—the Indian *maya*—as in a gallery of mirrors.”

And most strikingly:

“In all the world, there is no kind of framework within which we can find consciousness in the plural; this is simply something we construct because of the temporal plurality of individuals, but it is a false construction.”

The ontological ground: the unseen seer

The Vedantic position is not merely that “consciousness is fundamental” in the way that a materialist might say “matter is fundamental.” It is a more radical claim: consciousness is the *ontological ground*—the ground of being itself—from which everything else, including the material world, appears.

The Brihadaranyaka Upanishad states the principle with stark clarity:

You cannot see the seer of seeing. You cannot hear the hearer of hearing. You cannot think the thinker of thinking. You cannot know the knower of knowing. This is your Self, the inner controller, the immortal. Everything else is of no use.

The observer, in this tradition, is the *unseen seer*: the witness that cannot itself be witnessed, the subject that can never become an object. Every attempt to observe the observer produces another observed object (a brain scan, a neural correlate, a report of subjective experience)—never the observer itself. The observer is that which remains when every observed thing is removed: the irreducible first person, the ground on which the entire edifice of experience rests.

This ground cannot be objectified. It cannot be measured by the Born rule, because it is that *for which* the Born rule produces outcomes. It is not an eigenvalue of any observable; it is the condition of possibility of there being eigenvalues at all.

Tat tvam asi: thou art that. The deepest statement of Advaita Vedanta is the identification: this ontological ground—this unseen seer, this consciousness that cannot be objectified—is not separate from the ground of the universe itself. The individual witness (Atman) *is* the universal ground (Brahman). Not “is like” or “is connected to” or “participates in”: *is*, in the sense of identity. The finite perspective that looks through a particular brain and experiences a particular set of Born outcomes is a finite aperture of the same ground that runs the game engine.

The Vedantic observer in the Fock-Born framework

Translated into the language of this book:

The ontological ground runs the Fock-Schrödinger layer. The game engine ($\dot{h}_t = -iHh_t$) evolves the hidden state in the Fock space. In the Vedantic reading, this game engine is not a mathematical abstraction running on nothing: it is an expression of the ontological ground—Brahman, the one reality. The Fock space, the Hamiltonian, the unitary evolution are the inner

structure of this ground, the way it organises itself. The ground is not “made of” Fock basis vectors; the Fock basis vectors are how the ground appears when described mathematically.

The observer is at the level of the ground. The observer is not a subsystem of the Fock space. It is not a brain, not a detector, not an arrangement of quarks and electrons. These are objects *within* the game—Fock basis vectors, Born renderings, projected superpositions. The observer is the ground itself, experiencing the game from a finite perspective.

The brain is an avatar: a device within the rendered world, an interface through which the ground observes its own output. Just as a video game character is not the player but the player’s interface within the game, the brain is not the observer but the observer’s interface within the Born rendering. The brain is made of atoms; the atoms are Fock basis vectors; the Fock basis vectors are elements of the game engine. The observer is not made of any of these. The observer is the ground that runs the engine and witnesses its output through the avatar.

The Born rule renders Maya. The Born rendering layer— $P(\lambda) = \|P_\lambda h_t\|^2$ —produces the classical world: definite outcomes, definite positions, definite temperatures, the entire manifold of experienced reality. In the Vedantic reading, this rendered classical world is *Maya*: not an illusion in the sense of being non-existent, but an appearance that is not the fundamental reality. The fundamental reality is the Fock hidden state h_t (or more precisely, the ontological ground of which h_t is the mathematical description). Maya is the Born rendering of this ground: the projection onto eigensubspaces, the squared lengths, the definite outcomes, the classical world of separate objects and separate observers.

The classical world is real in the sense that it is experienced. It is not real in the sense that it is independent of the rendering. Between renderings—between the Born measurements—the classical world does not exist: there are no definite positions, no definite outcomes, only h_t ticking silently in the Fock space. Maya is the screen, not the game engine.

Consciousness underlies matter, not the reverse. The materialist position says: matter is fundamental; consciousness emerges from sufficiently complex arrangements of matter (brains, neural networks, information processing). The Vedantic position says: consciousness—the ontological ground, the unseen seer—is fundamental; matter is what appears when this ground is rendered through the Born layer.

In the Fock-Born framework, the materialist position identifies the observer with a subsystem of the Fock space: a specific arrangement of Fock basis vectors (a brain) that, through some as-yet-unknown mechanism, generates subjective experience. The Vedantic position identifies the observer with the ground that runs the Fock space: the brain is within the game; consciousness is not.

The difference is the direction of explanation. The materialist says: Fock space $\rightarrow$ complex arrangement $\rightarrow$ consciousness. The Vedantin says: consciousness (the ground) $\rightarrow$ Fock space (the game engine) $\rightarrow$ Born rendering (Maya) $\rightarrow$ the appearance of matter, brains, and separate observers within the rendered world.

The multiplicity of observers. The multiplicity of observers—Alice and Bob, a billion humans, every sentient being—is, in the Vedantic reading, a multiplicity of *finite perspectives* of the one ground, not a multiplicity of consciousnesses. Each perspective is an avatar: a physical subsystem (a brain, a detector) through which the ground observes its own Born output. The degeneracy of the observable determines what each avatar sees (classical or quantum, deterministic or random). But the witness behind every avatar—the unseen seer, the subject that cannot be objectified—is the same ground.

Schrödinger again:

“Consciousness is a singular of which the plural is unknown. . . . There is only one thing, and what seems to be a plurality is merely a series of different aspects of this one thing, produced by a deception—the Indian *maya*.”

The “deception” is the Born rendering: the projection onto eigensubspaces that produces the appearance of separate objects, separate observers, separate experiences. The ground is one. The avatars are many. The Born rule is the mechanism of Maya.

47.2 Bohr’s discipline: the question that must not be answered

Bohr’s position is the opposite of Schrödinger’s—not in content but in method. Bohr does not answer the question “what is the observer?” He refuses to ask it.

For Bohr, physics is the science of what can be communicated in unambiguous classical language. The experimental setup is described classically. The quantum formalism is a tool for predicting the classical outcomes. The observer is whoever reads the record. “What is the observer?” is not a physics question; it is a question about consciousness, and consciousness is not the subject of physics.

Bohr’s discipline is to draw a sharp line between what the formalism says and what it does not say. The formalism says: given an observable and a state, the Born rule gives probabilities. The formalism does not say: what the observer is, why outcomes occur, what the wave function “really is.” The formalism is complete on its side of the line. The questions on the other side are legitimate but belong to philosophy.

This discipline is enormously productive. It is why the Copenhagen axioms are so clean: they say exactly what is needed and nothing more. It is why the Fock-Born framework of this book works: the observer is undefined, and every chapter computes without impediment.

The cost is the permanent silence on the question Schrödinger found most important. Bohr does not deny the question; he denies that physics can answer it.

47.3 Heisenberg’s Platonism: potentiality and mathematical form

Heisenberg occupies the middle ground. He agreed with Bohr that the wave function is not a physical field, but he went further than Bohr in granting it ontological status: the wave function represents a *potentiality*—an objective tendency for outcomes to occur, intermediate between the actual and the non-existent.

In *Physics and Philosophy* (1958):

“The probability function. . . . represents a tendency for events and our knowledge of events. The probability function can be connected with reality only if one essential condition is fulfilled: if a new measurement is made to determine a certain property of the system.”

Heisenberg was a Platonist. He believed that the mathematical structures of physics—the symmetry groups, the Hilbert space, the Hamiltonian—are more real than the phenomena they describe:

“The elementary particles in Plato’s *Timaeus* are finally not substance but mathematical forms. ‘All things are numbers’ is a sentence attributed to Pythagoras. The only things that exist are mathematical forms.”

For Heisenberg, the observer is the agent that converts potentiality into actuality: the act of measurement is the transition from the possible to the real. But he identifies the observer with the *irreversible act of amplification*—the physical process by which a quantum event is amplified to a macroscopic record—not with consciousness. Any system capable of irreversible amplification is

an observer: a Geiger counter, a photographic plate, a cloud chamber. Consciousness is sufficient but not necessary.

Heisenberg did not go as far as Schrödinger. He did not claim that consciousness is fundamental, that there is only one consciousness, or that Atman is Brahman. He remained within Western philosophy—Plato, Aristotle, Kant—and did not engage substantively with Indian thought. His metaphysics is a metaphysics of mathematical structure, not of consciousness.

In the Fock-Born framework: Heisenberg’s position is that the Fock space and the Hamiltonian are the Platonic forms (more real than the phenomena); the Born output is the actualisation of the potentiality; and the observer is any irreversible amplifier. This is closer to Bohr’s operational stance than to Schrödinger’s Vedantic one, but it grants the mathematical structures a stronger reality than Bohr would allow.

47.4 Three positions, one silence

	The wave function is...	The observer is...	Philosophical tradition
Bohr	A tool for computing probabilities	Undefined (precondition, not subject)	Pragmatism, complementarity
Heisenberg	A potentiality (objective tendency)	Any irreversible amplifier	Platonism (forms more real than phenomena)
Schrödinger	A description within the one consciousness	Consciousness itself: singular, fundamental	Advaita Vedanta (<i>tat tvam asi</i>)

They disagree on nearly everything. But they agree on one point: the observer is not a concept that quantum mechanics defines internally. Bohr says it is undefined and must remain so. Heisenberg says it is an irreversible amplifier—but irreversibility is thermodynamic, not quantum-mechanical. Schrödinger says it is consciousness—which is not a physical concept at all.

Three of the deepest thinkers in the history of physics, who built the theory, who fought over its interpretation for decades, who understood it more intimately than anyone before or since, agreed: the observer is outside the formalism. The silence of the Copenhagen axioms is not an oversight. It is the considered judgment of the founders.

The Fock-Born framework of this book takes Bohr’s disciplined position as its working axiom: the observer is undefined; the Born rule is postulated; the formalism computes probabilities without defining what an outcome is or who experiences it.

This book has shown that this axiom is sufficient. With nothing more than $\dot{h}_t = -iHh_t$ and $P(\lambda) = \|P_\lambda h_t\|^2$, we have derived the Standard Model’s particle content, the interactions, the bound states, the condensates, the temperature, decoherence, the classical world, the collective modes, and the structure of spacetime as a Born-layer rendering. The observer was needed at every step —someone had to choose the observable—but was never defined, and the physics was never impeded by the omission.

The question remains. What is the observer?

Bohr says: do not ask. Heisenberg says: any irreversible amplifier. Schrödinger says: the ontological ground itself—the unseen seer, the subject that cannot be objectified—experiencing its own Born rendering through finite avatars. The brain is not the observer but the observer’s interface within the game. The classical world is not the fundamental reality but Maya: the Born rendering of the Fock hidden state. Consciousness does not emerge from the material

world; the material world appears within consciousness, rendered by the Born rule from the Fock hidden state that consciousness itself sustains. *Tat tvam asi.*

The Fock-Born framework does not choose. It cannot: the question is outside its scope. But it clarifies what the question is. Not “what collapses the wave function?” (the Born rule is a postulate, not a mechanism). Not “why are there outcomes?” (the Born rule says there are; it does not say why). The question is:

What is the relationship between the mathematical structure $P(\lambda) = \|P_\lambda h_t\|^2$ and the subjective experience of a definite outcome λ ?

This is, as Bohr recognized, not a physics question. It may be, as Schrödinger believed, the deepest question there is—or rather, not a question at all but the answer that precedes all questions: the ground on which the game engine runs, the unseen seer for whom the Born rule renders, the one reality of which the Fock hidden state is the mathematical shadow and the classical world is the projected screen.

The game engine ticks. The Born rule renders. The unseen seer witnesses. The question of what the seer is cannot be answered within the game—because the seer is not an object within the game but the ground on which the game is played.