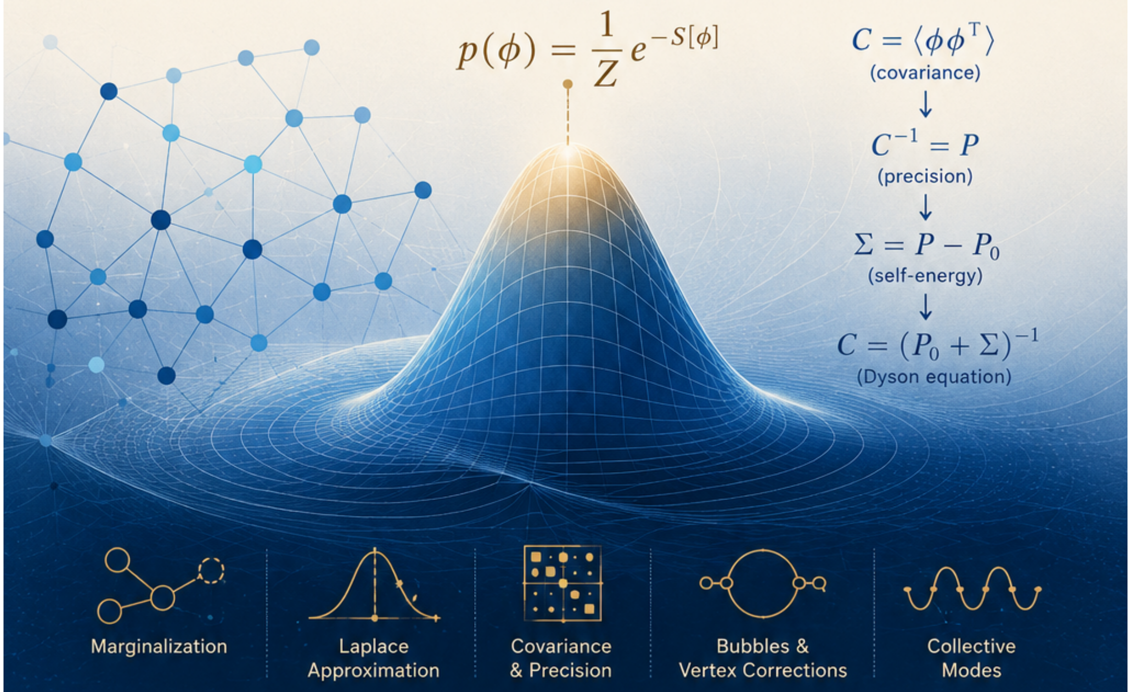
FIELD THEORY *from* COVARIANCE

A PROBABILISTIC INTRODUCTION TO QUANTUM FIELDS, COLLECTIVE MODES, AND MANY-BODY PHYSICS



A textbook for readers trained in probability, statistics, and machine learning

YING NIAN WU

UCLA

DRAFT EDITION

Field Theory from Covariance

A Probabilistic Introduction to
Quantum Fields, Collective Modes, and Many-Body Physics

A textbook for readers trained in probability, statistics, and machine learning

Ying Nian Wu, UCLA
written with the help of Claude 4.6 (book cover by GPT 5.5)

Contents

Preface

This book grew out of a simple observation: a large portion of what physicists compute in quantum field theory and many-body physics consists of operations that any well-trained statistician or machine-learning researcher already knows how to perform. These operations are marginalization, differentiation, matrix inversion, evaluation of log-determinants, and local Gaussian approximation. The difficulty, for someone entering from statistics or machine learning, is not the mathematics itself but the thick layer of specialized vocabulary, historical convention, and physical storytelling that surrounds it.

The traditional physics curriculum teaches field theory through a sequence of increasingly sophisticated physical pictures—particles scattering off one another, virtual processes mediating forces, quasiparticles dressing themselves with clouds of excitations, gauge bosons eating Goldstone bosons. These pictures are vivid and, once properly understood, genuinely illuminating. But they are not derivations. They are interpretations of mathematical structures, and the mathematical structures themselves are often simpler than the stories suggest.

This book takes the opposite approach. We begin with probability distributions, covariance matrices, precision matrices, cumulants, latent variables, and marginal likelihoods. We derive every result in finite-dimensional linear algebra before promoting it to functional notation. We name each mathematical object in the language of statistics and machine learning first, and introduce the physics terminology only after the derivation is complete. Physical intuition and storytelling come last—not because they are unimportant, but because they are most useful when the reader can clearly distinguish the story from the structure it illustrates.

The central object of study is a non-Gaussian probability distribution over a high-dimensional (or infinite-dimensional) random variable:

$$p(\phi) = \frac{1}{Z}\, e^{-S[\phi]}.$$

Here $S[\phi]$ is the negative log-density (physicists call it the *action*), and Z is the normalizing constant (physicists call it the *partition function*). The questions we ask about this distribution—its mean, its covariance, its conditional structure, its marginals, its response to perturbation, its spectral properties—are the same questions that arise throughout statistics. The answers, and the approximation methods used to obtain them, constitute a large portion of what is conventionally taught as quantum field theory.

We will see that *covariance* is what physicists call a *propagator*. *Precision*—the inverse of covariance—is the *inverse propagator*. The difference between the exact precision and a reference precision is the *self-energy*. The identity relating the exact covariance to the reference covariance and the self-energy is what physicists call the *Dyson equation*, and it is simply the statement that inverting the corrected precision gives the corrected covariance. Its infinite-series expansion is a Neumann series. The Laplace (saddle-point) approximation for collective-field fluctuations is what physicists call the *random phase approximation* (RPA). And a zero eigenvalue of a Hessian, arising from symmetry, is what physicists call a *Goldstone mode*.

None of this is intended to diminish the achievements of the physicists who discovered these structures. The physical problems that motivated these developments—superconductivity, the electron gas, quantum electrodynamics, phase transitions—are genuine triumphs of twentieth-century science, and the physical intuition that guided their solution is remarkable. Our goal is only to place that intuition in its correct logical position: after the mathematics, not in place of it.

A word about scope. This book treats *Euclidean* field theory, meaning that we work with statistical weights e^{-S} rather than quantum amplitudes e^{iS}. When the weight is positive and normalizable, it defines a genuine probability distribution, and all of our statistical language applies without qualification. Real-time quantum mechanics, unitary evolution, and the Born rule are discussed in the final part of the book, where we carefully explain what the Euclidean formulation computes, what it does not compute, and how the two are related through analytic continuation and spectral representations. We also note explicitly where the probability interpretation breaks down: Grassmann variables, fermion sign problems, gauge constraints, and complex actions all require additional care.

The intended reader has a working knowledge of probability, mathematical statistics, linear algebra, and ideally some exposure to machine learning, graphical models, or variational inference. No prior knowledge of physics is assumed, though readers with physics training may find the translation dictionary useful in the other direction.

A word about functional integrals. Throughout the book we write expressions such as $\int \mathcal{D}\phi \, e^{-S[\phi]}$, where ϕ is a field (a function of a continuous variable x). This notation should always be understood as a shorthand for the limit of a finite-dimensional integral on a lattice: ϕ is discretized to a vector ϕ_i on a lattice of N sites, the functional integral becomes $\int \prod_i d\phi_i \, e^{-S(\phi)}$, and the "continuum limit" is taken by sending the lattice spacing $a \to 0$ with appropriate adjustments of parameters (renormalization). We will make this precise in Part XI. Until then, the reader should interpret every functional integral as a finite-dimensional integral with n very large; the algebraic identities (completing the square, change of variables, Gaussian integration) are the same as in finite dimensions.

A Translation Dictionary

The following dictionary relates the terminology of probability, statistics, and linear algebra to the conventional vocabulary of physics. It appears here for reference, and an expanded version appears again at the end of the book. The reader should not attempt to memorize this table now; each entry will be derived and explained in the chapters that follow.

Probability, statistics, and linear algebra	**Physics terminology**
Random field	Euclidean field
Negative log-density	Action
Normalizing constant	Partition function
Log-normalizer	Free energy / connected generating functional
Covariance	Propagator
Precision (inverse covariance)	Inverse propagator
Reference covariance	Bare propagator
Exact covariance	Full or dressed propagator
Precision difference (exact minus reference)	Self-energy
Latent-variable augmentation	Hubbard–Stratonovich transformation
Negative log marginal density	Effective action
Mode of a marginal density	Saddle / mean-field configuration
Stationarity condition for the mode	Mean-field equation / gap equation
Mass scale induced at a saddle	Gap
Hessian of marginal negative log-density	Inverse collective propagator
Inverse Hessian of marginal negative log-density	Dressed collective propagator
Covariance of a composite observable	Susceptibility / polarization
Wick-factorized composite covariance	Bubble
Neumann expansion of corrected precision inverse	Dyson resummation
Quadratic (Laplace) approximation for collective fluctuations	RPA / Gaussian fluctuation theory
Null Hessian direction from symmetry	Goldstone mode
Small precision eigenvalue	Soft mode
Slow direction from a conservation law	Hydrodynamic mode
Pole of covariance	Collective excitation
Symmetry derivative identity	Ward identity
Higher derivatives of marginal negative log-density	Effective vertices
Non-Gaussian correction to simple bubble	Vertex correction

Probability, statistics, and linear algebra	**Physics terminology**
Marginalization over fast variables	Coarse-graining
Multiscale marginalization and rescaling	Renormalization

Part I

Euclidean Fields as Non-Gaussian Probability Models

Chapter 1

Euclidean Fields as Random Variables

1.1 The Setup

Consider a random variable ϕ taking values in $\mathbb{R}^n$. Suppose its probability density has the form

$$p(\phi) = \frac{1}{Z}\, e^{-S(\phi)}, \tag{1.1}$$

where $S\colon \mathbb{R}^n \to \mathbb{R}$ is a smooth function bounded below, and

$$Z = \int_{\mathbb{R}^n} e^{-S(\phi)}\, \mathrm{d}\phi \tag{1.2}$$

is the normalizing constant ensuring that p integrates to one. We assume throughout that S grows sufficiently fast at infinity to make Z finite.

This setup is entirely standard in statistics. The function $S(\phi)$ is the negative log-density of ϕ, up to the additive constant $\log Z$. In the language of exponential families, S specifies the natural parameters and sufficient statistics of the model, and $\log Z$ is the log-partition function.

In physics, S is called the **action**, and Z is called the **partition function**. We will use both vocabularies interchangeably, but the reader should always remember the underlying mathematical object: S is a function that determines a probability density through exponentiation and normalization. Nothing more, nothing less.

When n is large—and eventually we will consider the limit where ϕ becomes a function $\phi(x)$ defined on a continuous space, so that n is effectively infinite—the density $p(\phi)$ is concentrated in a complicated region of $\mathbb{R}^n$, and computing expectations under p becomes the central challenge. Almost everything in this book concerns methods for computing or approximating such expectations.

1.2 Moments and Cumulants

The mean of ϕ under p is

$$\langle \phi \rangle = \mathbb{E}[\phi] = \int \phi\, p(\phi)\, \mathrm{d}\phi.$$

The covariance matrix of ϕ is

$$C_{ij} = \mathrm{Cov}(\phi_i, \phi_j) = \langle \phi_i \phi_j \rangle - \langle \phi_i \rangle \langle \phi_j \rangle.$$

More generally, we can consider the k-th moment tensor $\langle \phi_{i_1} \cdots \phi_{i_k} \rangle$ and the k-th cumulant tensor. The relationship between moments and cumulants is the same as in ordinary statistics: the first cumulant is the mean, the second cumulant is the covariance, the third cumulant measures skewness, and so on. Cumulants of order three and higher vanish for Gaussian distributions, so they measure non-Gaussianity.

Physicists call moments **correlation functions** and cumulants **connected correlation functions**. The word “connected” refers to a diagrammatic property that we will encounter later: when moments are expanded in terms of Gaussian building blocks, cumulants correspond to contributions that cannot be factored into independent pieces. But the reader trained in statistics already knows this—it is the definition of a cumulant.

1.3 Source Tilting and the Generating Functional

A powerful technique for computing moments and cumulants is to introduce an auxiliary parameter—a source or tilting vector $J \in \mathbb{R}^n$—and define the **tilted partition function**

$$Z[J] = \int e^{-S(\phi)+J^\top \phi}\, \mathrm{d}\phi. \tag{1.3}$$

The term $J^\top \phi = \sum_i J_i \phi_i$ is an exponential tilt of the original density. In the language of exponential families, J plays the role of an additional natural parameter conjugate to ϕ. In physics, J is called an **external source**.

The **log-partition function** under the tilt is

$$W[J] = \log Z[J]. \tag{1.4}$$

This is the cumulant generating function of ϕ, evaluated at J (more precisely, it is the cumulant generating function of ϕ under the measure p, shifted by the constant $\log Z[0] = \log Z$). Physicists call $W[J]$ the **generating functional of connected correlation functions**, or sometimes simply the **free energy** (with a sign convention that varies across subfields).

The utility of $W[J]$ lies in differentiation. The first derivative of W with respect to J_i, evaluated at $J = 0$, gives the mean:

$$\left.\frac{\partial W}{\partial J_i}\right|_{J=0} = \frac{1}{Z}\int \phi_i\, e^{-S(\phi)}\, \mathrm{d}\phi = \langle \phi_i \rangle.$$

Let us derive this carefully. Starting from $W[J] = \log Z[J]$, we have

$$\frac{\partial W}{\partial J_i} = \frac{1}{Z[J]}\frac{\partial Z[J]}{\partial J_i}.$$

Now, $\frac{\partial}{\partial J_i} e^{-S(\phi)+J^\top\phi} = \phi_i\, e^{-S(\phi)+J^\top\phi}$. Therefore

$$\frac{\partial Z[J]}{\partial J_i} = \int \phi_i\, e^{-S(\phi)+J^\top\phi}\, \mathrm{d}\phi.$$

Combining these two lines:

$$\frac{\partial W}{\partial J_i} = \frac{\int \phi_i\, e^{-S(\phi)+J^\top\phi}\, \mathrm{d}\phi}{\int e^{-S(\phi)+J^\top\phi}\, \mathrm{d}\phi} = \mathbb{E}_J[\phi_i],$$

where $\mathbb{E}_J$ denotes expectation under the tilted density $p_J(\phi) \propto e^{-S(\phi)+J^\top\phi}$. At $J = 0$, this reduces to $\langle \phi_i \rangle$. In vector notation:

$$\nabla_J W[J] = \mathbb{E}_J[\phi]. \tag{1.5}$$

The second derivative gives the covariance under the tilted measure. Let us derive this as well. Taking a further derivative:

$$\frac{\partial^2 W}{\partial J_i\,\partial J_j} = \frac{\partial}{\partial J_j}\mathbb{E}_J[\phi_i].$$

To evaluate this, we use the identity

$$\frac{\partial}{\partial J_j}\mathbb{E}_J[\phi_i] = \mathbb{E}_J[\phi_i\phi_j] - \mathbb{E}_J[\phi_i]\,\mathbb{E}_J[\phi_j],$$

which is a standard result in the theory of exponential families (it expresses the fact that the second derivative of the log-partition function is the variance). For completeness, here is the derivation. Write $\mathbb{E}_J[\phi_i] = \frac{\int \phi_i\, e^{-S+J^\top\phi}\,\mathrm{d}\phi}{Z[J]}$. Differentiating with respect to J_j by the quotient rule:

$$\frac{\partial}{\partial J_j}\mathbb{E}_J[\phi_i] = \frac{\int \phi_i\phi_j\, e^{-S+J^\top\phi}\,\mathrm{d}\phi}{Z[J]} - \frac{\int \phi_i\, e^{-S+J^\top\phi}\,\mathrm{d}\phi}{Z[J]}\cdot\frac{\int \phi_j\, e^{-S+J^\top\phi}\,\mathrm{d}\phi}{Z[J]}.$$

The first term is $\mathbb{E}_J[\phi_i\phi_j]$ and the second is $\mathbb{E}_J[\phi_i]\,\mathbb{E}_J[\phi_j]$. Therefore

$$\frac{\partial^2 W}{\partial J_i\,\partial J_j} = \mathrm{Cov}_J(\phi_i, \phi_j). \tag{1.6}$$

In matrix notation:

$$\nabla_J^2 W[J] = \mathrm{Cov}_J(\phi). \tag{1.7}$$

At $J = 0$, this is simply the covariance matrix $C = \mathrm{Cov}(\phi)$.

This pattern continues: the k-th derivative of W with respect to J gives the k-th cumulant of ϕ under the tilted measure. This is why physicists call W the generator of connected correlation functions—derivatives of W produce cumulants, which are the "connected" parts of the moments.

1.4 Why Covariance Is Central

Among all the statistical properties of ϕ, the covariance matrix C occupies a special position. It tells us which components of ϕ fluctuate together, how large those fluctuations are, and—through its eigendecomposition—what the natural coordinates of the distribution are.

For Gaussian distributions, the covariance (together with the mean) determines the entire distribution. For non-Gaussian distributions, the covariance captures only part of the story, but it remains the most tractable two-point summary. Much of this book is concerned with computing or approximating the exact covariance of non-Gaussian distributions, and with understanding how that covariance differs from the covariance of a simpler Gaussian reference.

We will find that the difference between the exact precision (inverse covariance) and a reference precision is a quantity of enormous importance. Physicists call it the *self-energy*, and it encodes the effects of non-Gaussianity—interactions, in physics language—on the two-point correlation structure. But before we can appreciate what self-energy means, we need to develop the Gaussian case thoroughly.

1.5 Looking Ahead

In the coming chapters, we will develop the Gaussian case in detail (Chapter 2), then study what happens when the distribution is non-Gaussian and we examine composite observables like ϕ_i^2 (Chapter 3). Part II will introduce a powerful technique—latent-variable augmentation—that converts certain non-Gaussian problems into conditionally Gaussian ones. Part III will show how marginalization over Gaussian variables produces log-determinants, a structure that pervades all of field theory.

Throughout, we will work first in finite dimensions, with explicit matrices and vectors, before moving to the continuum limit where ϕ becomes a field $\phi(x)$ and matrices become integral kernels. The reader should always be able to return to the finite-dimensional picture for clarity.

Chapter 2

Gaussian Fields, Covariance, and Precision

2.1 The Gaussian Density

The simplest member of the family (1.1) is the Gaussian distribution, obtained when $S(\phi)$ is a quadratic function. Let A be a real, symmetric, positive-definite $n \times n$ matrix. The density

$$p_0(\phi) = \frac{1}{Z_0} \exp\left(-\frac{1}{2} \phi^\top A \phi \right), \tag{2.1}$$

with Z_0 chosen to normalize, defines a zero-mean Gaussian distribution. (We write p_0 and Z_0 with a subscript to indicate that this is a reference or "bare" distribution, to which we will later compare the full non-Gaussian measure.)

The matrix A is the **precision matrix** of this Gaussian—it is the inverse of the covariance matrix. To derive this fundamental fact, we need the Gaussian integral.

2.2 The Gaussian Integral

We wish to evaluate

$$Z_0 = \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} x^\top A x \right) \, \mathrm{d}x. \tag{2.2}$$

The idea is to reduce this to a product of one-dimensional integrals by diagonalizing A. Since A is real and symmetric, there exists an orthogonal matrix O and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i > 0$ such that $A = O \Lambda O^\top$. Making the change of variables $y = O^\top x$ (which has Jacobian $|\det O| = 1$, since O is orthogonal), we get

$$x^\top A x = y^\top \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2,$$

and therefore

$$Z_0 = \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \lambda_i y_i^2 \right) \, \mathrm{d}y = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \lambda_i y_i^2 \right) \, \mathrm{d}y_i.$$

Each factor is a standard one-dimensional Gaussian integral. Recall that $\int_{-\infty}^{\infty} e^{-\alpha t^2}\,\mathrm{d}t = \sqrt{\pi/\alpha}$ for $\alpha > 0$. Setting $\alpha = \lambda_i/2$:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\lambda_i y_i^2\right)\,\mathrm{d}y_i = \sqrt{\frac{2\pi}{\lambda_i}}.$$

Multiplying:

$$Z_0 = \prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda_i}} = (2\pi)^{n/2}\left(\prod_{i=1}^{n}\lambda_i\right)^{-1/2} = (2\pi)^{n/2}(\det A)^{-1/2}. \tag{2.3}$$

This is the n-dimensional Gaussian integral. It tells us that the normalizing constant depends on the determinant of the precision matrix—a fact that will become important when A itself depends on parameters.

2.3 Covariance Equals the Inverse Precision

To find the covariance matrix of ϕ under p_0, we use the source-tilting technique from Chapter 1. Define

$$Z_0[J] = \int \exp\left(-\frac{1}{2}\phi^\top A\phi + J^\top\phi\right)\,\mathrm{d}\phi.$$

We evaluate this by **completing the square**. The exponent is

$$-\frac{1}{2}\phi^\top A\phi + J^\top\phi.$$

We want to write this in the form $-\frac{1}{2}(\phi - \text{something})^\top A(\phi - \text{something}) + \text{constant}$. Let us try $\phi - A^{-1}J$. Expanding:

$$-\frac{1}{2}(\phi - A^{-1}J)^\top A(\phi - A^{-1}J) = -\frac{1}{2}\phi^\top A\phi + \phi^\top AA^{-1}J - \frac{1}{2}(A^{-1}J)^\top A(A^{-1}J).$$

The middle term simplifies: $\phi^\top AA^{-1}J = \phi^\top J = J^\top\phi$. The last term: $(A^{-1}J)^\top A(A^{-1}J) = J^\top A^{-1}AA^{-1}J = J^\top A^{-1}J$. Therefore

$$-\frac{1}{2}\phi^\top A\phi + J^\top\phi = -\frac{1}{2}(\phi - A^{-1}J)^\top A(\phi - A^{-1}J) + \frac{1}{2}J^\top A^{-1}J. \tag{2.4}$$

This is exactly the technique of completing the square, expressed in matrix notation. The shift $\phi \to \phi - A^{-1}J$ does not change the integration measure (it is a translation), so

$$Z_0[J] = \exp\left(\frac{1}{2}J^\top A^{-1}J\right)\int \exp\left(-\frac{1}{2}\eta^\top A\eta\right)\,\mathrm{d}\eta = Z_0\cdot\exp\left(\frac{1}{2}J^\top A^{-1}J\right).$$

Taking the logarithm:

$$W_0[J] = \log Z_0[J] = \log Z_0 + \frac{1}{2}J^\top A^{-1}J. \tag{2.5}$$

Now we read off the cumulants. The first derivative:

$$\frac{\partial W_0}{\partial J_i} = (A^{-1}J)_i.$$

At $J = 0$, this gives $\langle\phi_i\rangle = 0$, confirming the zero mean. The second derivative:

$$\frac{\partial^2 W_0}{\partial J_i\,\partial J_j} = (A^{-1})_{ij}.$$

This is independent of J, and it gives us the covariance:

$$\boxed{C_0 = \mathrm{Cov}(\phi) = A^{-1}.} \tag{2.6}$$

This is the fundamental relationship: **the covariance matrix of a Gaussian distribution is the inverse of its precision matrix**. In the physics vocabulary, C_0 is the **bare propagator** and A is the **inverse bare propagator**.

Since $W_0[J]$ is exactly quadratic in J, all cumulants of order three and higher vanish. This is the defining property of Gaussianity: a Gaussian distribution is completely determined by its first two cumulants (mean and covariance).

2.4 Precision and Covariance: Two Descriptions of the Same Distribution

It is worth pausing to appreciate the dual roles of A and $C_0 = A^{-1}$.

The precision matrix A appears naturally in the exponent of the density. It describes the *cost of fluctuations*: if A_{ii} is large, then ϕ_i is tightly constrained near zero, because the quadratic penalty $\frac{1}{2}A_{ii}\phi_i^2$ is steep. Off-diagonal entries A_{ij} with $i \neq j$ create coupling between components—the cost of ϕ_i taking a particular value depends on the value of ϕ_j.

The covariance matrix $C_0 = A^{-1}$, on the other hand, describes the *statistics of the resulting fluctuations*. Even though A may be sparse (so that the direct couplings are local), the covariance C_0 is generally dense—fluctuations in ϕ_i can be correlated with fluctuations in ϕ_j even if $A_{ij} = 0$, because the correlation propagates through chains of nonzero entries in A. This phenomenon—local precision producing long-range covariance—is central to the physics of propagation.

In graphical models, the precision matrix encodes the **conditional independence structure**: ϕ_i and ϕ_j are conditionally independent given all other components if and only if $A_{ij} = 0$. The covariance matrix, in contrast, encodes the **marginal correlation structure**. Inverting the precision to get the covariance translates conditional structure into marginal structure, and this translation can produce correlations between variables that are conditionally independent.

This is precisely what physicists describe when they say that a particle *propagates* through a medium. The precision matrix specifies the local interactions (the medium), and the covariance—the propagator—describes how disturbances in one location correlate with disturbances in another. The propagator is obtained by inverting the precision.

2.5 Gaussian Conditioning and the Schur Complement

Suppose we partition ϕ into two groups, $\phi = (\phi_a, \phi_b)$, and correspondingly partition the precision matrix:

$$A = \begin{pmatrix} A_{aa} & A_{ab} \\ A_{ba} & A_{bb} \end{pmatrix}.$$

What is the conditional distribution of ϕ_a given ϕ_b? And what is the marginal distribution of ϕ_a?

For a Gaussian, both questions have clean answers expressed in terms of Schur complements.

The **conditional distribution** of ϕ_a given $\phi_b = \beta$ is Gaussian with precision matrix A_{aa} and mean $\mu_{a|b} = -A_{aa}^{-1}A_{ab}\beta$. That is:

$$p(\phi_a \mid \phi_b = \beta) \propto \exp\left(-\frac{1}{2}(\phi_a - \mu_{a|b})^\top A_{aa}(\phi_a - \mu_{a|b})\right).$$

Let us derive this. The joint density involves $\frac{1}{2}\phi^\top A\phi$. Expanding in blocks:

$$\phi^\top A\phi = \phi_a^\top A_{aa}\phi_a + 2\phi_a^\top A_{ab}\phi_b + \phi_b^\top A_{bb}\phi_b.$$

Conditioning on $\phi_b = \beta$ means treating β as fixed. The terms involving only β become constants (absorbed into the normalization), and we are left with

$$-\frac{1}{2}\phi_a^\top A_{aa}\phi_a - \phi_a^\top A_{ab}\beta + \text{const.}$$

This is a quadratic in ϕ_a, so the conditional is Gaussian with precision A_{aa} and a linear shift determined by $A_{ab}\beta$. Completing the square (exactly as in Section 2.3):

$$-\frac{1}{2}\phi_a^\top A_{aa}\phi_a - \phi_a^\top A_{ab}\beta = -\frac{1}{2}(\phi_a + A_{aa}^{-1}A_{ab}\beta)^\top A_{aa}(\phi_a + A_{aa}^{-1}A_{ab}\beta) + \text{const.}$$

The conditional mean is $\mu_{a|b} = -A_{aa}^{-1}A_{ab}\beta$ and the conditional covariance is A_{aa}^{-1}.

Notice a remarkable fact: the conditional precision of ϕ_a given ϕ_b is simply A_{aa}, the upper-left block of the joint precision matrix. The off-diagonal blocks A_{ab} affect only the conditional mean, not the conditional precision. This is a special property of the Gaussian.

The **marginal distribution** of ϕ_a is obtained by integrating out ϕ_b. The marginal covariance of ϕ_a is the upper-left block of the joint covariance $C_0 = A^{-1}$, which can be computed by the Schur complement formula. We need to compute

$$\int \exp\left(-\frac{1}{2}\phi^\top A\phi\right)\,\mathrm{d}\phi_b.$$

Expanding the quadratic in blocks and treating ϕ_a as fixed:

$$-\frac{1}{2}\phi_b^\top A_{bb}\phi_b - \phi_b^\top A_{ba}\phi_a - \frac{1}{2}\phi_a^\top A_{aa}\phi_a.$$

The integral over ϕ_b is a Gaussian integral with precision A_{bb} and source $-A_{ba}\phi_a$. By the completing-the-square result (2.4):

$$\int \exp\left(-\frac{1}{2}\phi_b^\top A_{bb}\phi_b - \phi_b^\top A_{ba}\phi_a\right)\,\mathrm{d}\phi_b \propto \exp\left(\frac{1}{2}\phi_a^\top A_{ab}A_{bb}^{-1}A_{ba}\phi_a\right).$$

Combining with the remaining $-\frac{1}{2}\phi_a^\top A_{aa}\phi_a$ term:

$$p(\phi_a) \propto \exp\left(-\frac{1}{2}\phi_a^\top \underbrace{\left(A_{aa} - A_{ab}A_{bb}^{-1}A_{ba}\right)}_{\text{Schur complement}}\phi_a\right). \tag{2.7}$$

The marginal precision of ϕ_a is the **Schur complement** $A_{aa} - A_{ab}A_{bb}^{-1}A_{ba}$. It is lower than the conditional precision A_{aa}, reflecting the additional uncertainty introduced by not knowing ϕ_b. In a precise sense: integrating out variables reduces the effective precision and increases the effective covariance.

This is a pattern that will recur throughout the book. When we integrate out (marginalize over) some variables, the effective precision of the remaining variables changes. The change in precision—the difference between what we had before and what we have after—encodes the effect of the integrated-out degrees of freedom. In later chapters, this will be the origin of the self-energy.

2.6 Wick's Theorem

Wick's theorem is the statement that all moments of a Gaussian distribution can be expressed in terms of the covariance. For a zero-mean Gaussian with covariance C_0:

$$\langle \phi_{i_1} \phi_{i_2} \cdots \phi_{i_{2k}} \rangle = \sum_{\text{pairings}} \prod_{\text{pairs } (a,b)} C_0(i_a, i_b), \tag{2.8}$$

and all odd moments vanish. The sum runs over all ways of partitioning the $2k$ indices into k pairs. The number of such pairings is $(2k-1)!! = (2k-1)(2k-3)\cdots 3 \cdot 1$.

Let us verify this for the simplest nontrivial case: the fourth moment. For a zero-mean Gaussian:

$$\langle \phi_i \phi_j \phi_k \phi_l \rangle = C_0(i,j)\, C_0(k,l) + C_0(i,k)\, C_0(j,l) + C_0(i,l)\, C_0(j,k). \tag{2.9}$$

This can be derived by differentiating $W_0[J] = \frac{1}{2} J^\top C_0 J$ four times with respect to J, or by explicit computation using the eigendecomposition of C_0.

Wick's theorem is the workhorse of perturbation theory. It tells us that when we compute expectations of products of Gaussian random variables, the answer decomposes into products of pairwise covariances. Each "pairing" corresponds to what physicists call a **contraction**, and the diagrammatic representation of Wick's theorem is the origin of Feynman diagrams.

Derivation of Wick's theorem for the fourth moment

Let us verify equation (2.9) by explicit computation using the generating function. Recall that for a zero-mean Gaussian with covariance C_0, the generating function is $W_0[J] = \frac{1}{2} J^\top C_0 J$, and the moment generating function is $Z_0[J] = Z_0 \exp(\frac{1}{2} J^\top C_0 J)$. The fourth moment is

$$\langle \phi_i \phi_j \phi_k \phi_l \rangle = \frac{\partial^4}{\partial J_i\, \partial J_j\, \partial J_k\, \partial J_l} Z_0[J] \bigg|_{J=0} \cdot \frac{1}{Z_0}.$$

Write $f(J) = \exp(\frac{1}{2} J^\top C_0 J)$. The first derivative is $\frac{\partial f}{\partial J_i} = (C_0 J)_i\, f$. Taking a second derivative:

$$\frac{\partial^2 f}{\partial J_i \partial J_j} = [C_0(i,j) + (C_0 J)_i (C_0 J)_j]\, f.$$

Taking a third:

$$\frac{\partial^3 f}{\partial J_i \partial J_j \partial J_k} = [C_0(i,j)(C_0 J)_k + C_0(i,k)(C_0 J)_j + C_0(j,k)(C_0 J)_i + (C_0 J)_i (C_0 J)_j (C_0 J)_k]\, f.$$

At $J = 0$, every term with a factor of $(C_0 J)$ vanishes, and $f(0) = 1$, so the third moment is zero (as expected for a zero-mean Gaussian). The fourth derivative at $J = 0$ collects only the terms where every $(C_0 J)$ factor has been differentiated, giving exactly the three pairings:

$$\frac{\partial^4 f}{\partial J_i \partial J_j \partial J_k \partial J_l} \bigg|_{J=0} = C_0(i,j) C_0(k,l) + C_0(i,k) C_0(j,l) + C_0(i,l) C_0(j,k).$$

The general result follows by induction: the $2k$-th derivative of $\exp(\frac{1}{2} J^\top C_0 J)$ at $J = 0$ produces the sum over all $(2k-1)!!$ pairings.

2.7 Fourier Diagonalization of Translation-Invariant Covariances

In many applications, the random variable ϕ is indexed by a spatial location: $\phi_i \to \phi(x)$, where x ranges over a lattice or a continuous space. When the precision matrix (or kernel) depends only on the difference $x - y$—that is, when the distribution is translation-invariant—the covariance can be diagonalized by a Fourier transform.

Consider a lattice of N sites with periodic boundary conditions. A translation-invariant precision matrix $A(x, y) = A(x - y)$ is diagonalized by the discrete Fourier transform:

$$\tilde{A}(k) = \sum_r A(r)\, e^{-ik\cdot r}.$$

The covariance in Fourier space is then simply

$$\tilde{C}_0(k) = \frac{1}{\tilde{A}(k)}, \tag{2.10}$$

because matrix inversion of a circulant matrix reduces to pointwise inversion in Fourier space.

For example, the lattice Laplacian $A(x, y) = -\Delta + m_0^2$ has Fourier transform $\tilde{A}(k) = \sum_\mu 2(1 - \cos k_\mu) + m_0^2 \approx k^2 + m_0^2$ for small k. The covariance is then

$$\tilde{C}_0(k) = \frac{1}{k^2 + m_0^2}.$$

In real space, this corresponds to a covariance that decays exponentially with distance, with a correlation length $\xi = 1/m_0$. When $m_0 = 0$, the covariance decays as a power law. These are facts about the Fourier transform of $1/(k^2 + m_0^2)$, but physicists describe them using more evocative language: a nonzero m_0 gives the field a **mass**, and the correlation length is the inverse mass. We will return to this at length in Part IV.

Worked example: Fourier diagonalization on a chain

To make this concrete, consider a one-dimensional chain of $N = 8$ sites with periodic boundary conditions. The nearest-neighbor precision matrix has entries $A_{ij} = (2 + m_0^2)\delta_{ij} - \delta_{|i-j|,1} - \delta_{|i-j|,N-1}$, where the last term enforces periodicity. This is a circulant matrix: each row is a cyclic shift of the previous one.

Circulant matrices are diagonalized by the discrete Fourier transform (DFT) matrix $F_{jk} = \frac{1}{\sqrt{N}} e^{-2\pi \mathrm{i} jk/N}$. The eigenvalues are

$$\tilde{A}(k_n) = 2 + m_0^2 - 2\cos(2\pi n/N), \qquad n = 0, 1, \ldots, N-1,$$

where $k_n = 2\pi n/N$ is the lattice momentum. For small k_n, we have $2 - 2\cos k_n \approx k_n^2$, recovering $\tilde{A}(k) \approx k^2 + m_0^2$.

The covariance in Fourier space is $\tilde{C}_0(k_n) = 1/\tilde{A}(k_n)$, and the real-space covariance is

$$C_0(r) = \frac{1}{N} \sum_{n=0}^{N-1} \frac{e^{\mathrm{i} k_n r}}{2 + m_0^2 - 2\cos k_n}.$$

For concreteness, let $N = 8$ and $m_0^2 = 1$. Then the eigenvalues are $\tilde{A}(k_n) = 3 - 2\cos(\pi n/4)$ for $n = 0, \ldots, 7$: the values are $1, 3 - \sqrt{2}, 3, 3 + \sqrt{2}, 5, 3 + \sqrt{2}, 3, 3 - \sqrt{2}$. The covariance at separation r is then a sum of eight terms, each the reciprocal of an eigenvalue weighted by a phase.

For $m_0 > 0$, the real-space covariance $C_0(r)$ decays exponentially: $C_0(r) \sim e^{-m_0|r|}$ for $|r| \gg 1$, with correlation length $\xi = 1/m_0$. When $m_0 = 0$, the smallest eigenvalue is $\tilde{A}(0) = 0$ (for the zero mode), the matrix A is singular, and $C_0(r)$ decays as a power law—there is no finite correlation length. The reader should verify these claims numerically for the $N = 8$ example above: the calculation is a finite matrix inversion.

This concrete example illustrates the general pattern: a translation-invariant (circulant) precision matrix is diagonalized by the DFT, matrix inversion becomes pointwise division in Fourier space, and the decay properties of the covariance are controlled by the smallest eigenvalue of the precision—which is the eigenvalue at zero momentum.

Chapter 3

Non-Gaussianity and Composite Observables

3.1 Beyond Gaussianity

The Gaussian distribution is analytically tractable because all its cumulants of order three and higher vanish, and all moments factor into products of covariances via Wick's theorem. Real distributions of interest are almost never Gaussian. In physics, the non-Gaussian terms in the action $S(\phi)$ are called **interactions**, because they couple components of ϕ in ways that go beyond pairwise (quadratic) coupling.

Consider the simplest non-Gaussian generalization: a one-dimensional quartic density

$$p(x) \propto \exp\left(-\frac{1}{2}ax^2 - \frac{\lambda}{4}x^4\right), \tag{3.1}$$

with $a > 0$ and $\lambda > 0$. The quadratic term $\frac{1}{2}ax^2$ gives a Gaussian with precision a (variance $1/a$). The quartic term $\frac{\lambda}{4}x^4$ penalizes large fluctuations more severely than the Gaussian does, making the distribution more sharply peaked.

Even for this simple density, the exact covariance $\mathrm{Var}(x) = \langle x^2 \rangle$ cannot be computed in closed form (it involves ratios of parabolic cylinder functions). We will use it as a running example throughout the book to illustrate the various approximation methods—Laplace approximation, perturbative expansion, latent-variable augmentation—as they are introduced.

3.2 Composite Observables and Their Covariances

Suppose that instead of asking about ϕ itself, we are interested in some nonlinear function of ϕ. Let $O(\phi)$ be such a **composite observable**—for example, $O(\phi) = \phi(x)^2$, or $O(\phi) = \phi(x)\phi(y)$, or a more complicated nonlinear function.

The covariance of a composite observable is

$$\chi_O = \mathrm{Cov}(O, O) = \langle O^2 \rangle - \langle O \rangle^2.$$

This quantity—the variance of a nonlinear function of ϕ—is what physicists call a **susceptibility**. It measures how much the observable O fluctuates. In condensed-matter physics, susceptibility has a more specific meaning tied to the response of a system to an external perturbation, but the connection between fluctuation and response (the fluctuation-dissipation relation) means that the two uses are closely related. We will derive this connection explicitly in Chapter 16.

3.3 The Bubble: Gaussian Covariance of a Quadratic Observable

Let us compute the covariance of the composite observable $O(x) = \phi(x)^2$ under a zero-mean Gaussian distribution with covariance $C(x, y)$. This is the simplest example of a composite-observable covariance, and it introduces one of the most important structures in all of field theory.

We need to compute

$$\langle \phi(x)^2 \phi(y)^2 \rangle - \langle \phi(x)^2 \rangle \langle \phi(y)^2 \rangle.$$

The second term is straightforward. For a zero-mean Gaussian, $\langle \phi(x)^2 \rangle = C(x, x)$, so $\langle \phi(x)^2 \rangle \langle \phi(y)^2 \rangle = C(x, x)\, C(y, y)$.

The first term requires the fourth moment $\langle \phi(x)^2 \phi(y)^2 \rangle$. By Wick's theorem (2.9), with all four indices specialized appropriately, we label the four fields as $\phi_1 = \phi_2 = \phi(x)$ and $\phi_3 = \phi_4 = \phi(y)$. The three pairings give:

- Pair $(1, 2)$ and $(3, 4)$: $C(x, x) \cdot C(y, y)$.
- Pair $(1, 3)$ and $(2, 4)$: $C(x, y) \cdot C(x, y)$.
- Pair $(1, 4)$ and $(2, 3)$: $C(x, y) \cdot C(x, y)$.

Therefore

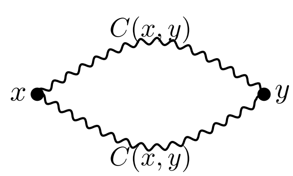
$$\langle \phi(x)^2 \phi(y)^2 \rangle = C(x, x)\, C(y, y) + 2\, C(x, y)^2.$$

Subtracting:

$$\boxed{\operatorname{Cov}\big(\phi(x)^2,\, \phi(y)^2\big) = 2\, C(x, y)^2.} \tag{3.2}$$

This is a beautiful and important result. The covariance of the composite observable ϕ^2 is a product of two copies of the fundamental covariance C.

Physicists draw this as a **bubble diagram**: two propagator lines running from x to y and from y to x, forming a closed loop (a "bubble"):



But the diagram is just a pictorial representation of the algebraic fact (3.2). The "bubble" is the Gaussian covariance of a composite quadratic observable. We can state this as a key principle:

> The bubble is the Gaussian covariance of a composite observable. It arises whenever we ask how much a quadratic function of Gaussian random variables fluctuates, and it equals a product of two covariance kernels, summed over internal indices.

The factor of 2 in (3.2) is a combinatorial factor counting the two nontrivial Wick pairings. In more general situations, the combinatorics become more complicated, but the underlying structure—products of covariance kernels contracted over internal indices—remains the same.

3.4 The Bubble in Fourier Space

For a translation-invariant covariance $C(x, y) = C(x-y)$, the Fourier transform of the bubble is particularly illuminating. Define $\Pi(q)$ as the Fourier transform of $\mathrm{Cov}(\phi(x)^2, \phi(y)^2)$ (up to normalizing conventions). In Fourier space:

$$\Pi(q) = 2 \int_k C(k+q)\, C(k), \tag{3.3}$$

where $\int_k$ denotes integration (or summation) over the internal momentum k. This convolution structure—two propagators sharing external momentum q—is the momentum-space bubble.

Notice that this is simply the Fourier representation of a product of two functions evaluated at the same pair of points, turned into a convolution by the convolution theorem. There is nothing mysterious about it, once we recognize that it started as a covariance of a composite observable.

Derivation of the Fourier-space bubble

Let us derive equation (3.3) from the real-space result (3.2). The Fourier transform of the bubble is defined as

$$\Pi(q) = \int \mathrm{d}^d r\, e^{-\mathrm{i}q\cdot r}\ \mathrm{Cov}(\phi(0)^2, \phi(r)^2) = 2 \int \mathrm{d}^d r\, e^{-\mathrm{i}q\cdot r}\, C(r)^2.$$

Now, $C(r) = \int \frac{d^d k}{(2\pi)^d}\, \tilde{C}(k)\, e^{\mathrm{i}k\cdot r}$. Therefore

$$C(r)^2 = \int \frac{d^d k_1}{(2\pi)^d} \frac{d^d k_2}{(2\pi)^d}\, \tilde{C}(k_1)\tilde{C}(k_2)\, e^{\mathrm{i}(k_1+k_2)\cdot r}.$$

Inserting into the Fourier transform and integrating over r produces $\delta(k_1 + k_2 - q)(2\pi)^d$:

$$\Pi(q) = 2 \int \frac{d^d k}{(2\pi)^d}\, \tilde{C}(k)\, \tilde{C}(q-k).$$

Relabeling $k \to k$ and using translation invariance ($\tilde{C}(q-k) = \tilde{C}(k+q)$ when C is even in momentum, which holds for real symmetric $C(r)$), we recover (3.3). This is the convolution theorem: the Fourier transform of a product of functions is the convolution of their Fourier transforms.

For the free propagator $\tilde{C}(k) = 1/(k^2 + m^2)$, the bubble integral is

$$\Pi(q) = 2 \int \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2 + m^2)((k+q)^2 + m^2)}.$$

This integral—two propagators sharing external momentum q—is the workhorse of quantum field theory. We will evaluate it explicitly in several dimensions in later chapters. For now, note two limiting behaviors. At $q = 0$: $\Pi(0) = 2 \int_k 1/(k^2 + m^2)^2$, which converges in $d < 4$ and diverges logarithmically in $d = 4$. At large $|q| \gg m$: $\Pi(q) \sim |q|^{d-4}$ (up to logarithms), which tells us that the bubble becomes a more severe divergence in higher dimensions.

3.5 Susceptibility and Response

The covariance of a composite observable is closely related to the concept of **response**—how much the expected value of O changes when we perturb the distribution. Suppose we add a source term $J \cdot O$ to the action:

$$S_J(\phi) = S(\phi) - J\, O(\phi).$$

Then $\langle O\rangle_J = \frac{\partial W}{\partial J}$, where $W[J] = \log \int e^{-S(\phi)+JO(\phi)}\,\mathrm{d}\phi$, and

$$\left.\frac{\partial \langle O\rangle_J}{\partial J}\right|_{J=0} = \mathrm{Cov}(O,O) = \chi_O. \tag{3.4}$$

This identity—the derivative of the expectation with respect to the source equals the covariance—is a special case of the general result (1.6). It tells us that the fluctuation of an observable (its variance) equals its linear response to the conjugate source. Physicists call this the **fluctuation-dissipation relation** (in its simplest, equilibrium form), and the quantity χ_O is the susceptibility.

3.6 Non-Gaussian Corrections

Under a non-Gaussian distribution, the covariance of ϕ^2 is no longer given by the simple formula (3.2). The Wick factorization fails, and there are additional "connected" contributions involving the non-Gaussian parts of the distribution. These corrections are called **vertex corrections** in physics, and they measure the departure of the composite covariance from its Gaussian (factorized) value.

We will return to this in detail in Part IX. For now, the key point is this: equation (3.2) gives the bubble as the Gaussian covariance of a composite observable, and this is the leading-order result in many approximation schemes. But it is not generally exact when the underlying distribution is non-Gaussian.

What was exact, what was approximate, and what was conventional?

The results in this chapter are organized as follows.

Exact identities. The definition of composite-observable covariance is exact. The relation between covariance and response (3.4) is exact. Wick's theorem (2.8) is exact for Gaussian distributions.

Results specific to the Gaussian. The bubble formula (3.2)—the statement that $\mathrm{Cov}(\phi^2(x), \phi^2(y)) = 2C(x,y)^2$—is exact for a zero-mean Gaussian distribution. It becomes an approximation when applied to non-Gaussian distributions.

Conventions. The term "bubble" refers to the diagrammatic representation of the product $C(x,y)^2$. The term "susceptibility" for χ_O is standard but context-dependent: the same mathematical object can be called susceptibility, response, polarization, or density-density correlation depending on what O represents.

Part II

Latent-Variable Augmentation

Chapter 4

Latent-Variable Augmentation

4.1 The Central Idea

Many non-Gaussian distributions become simpler—sometimes even Gaussian—when viewed as marginals of a higher-dimensional joint distribution. This is a familiar idea in statistics: latent-variable models, data augmentation, auxiliary-variable methods, and hierarchical Bayesian models all exploit the same principle. The key insight is:

> If $p(\phi) = \int p(\phi, \sigma)\, d\sigma$, and if $p(\phi \mid \sigma)$ or $p(\sigma \mid \phi)$ has a tractable form, then the complicated marginal $p(\phi)$ can be studied through the simpler conditional.

In physics, this principle is realized through the **Hubbard–Stratonovich transformation** (HST), which converts a non-Gaussian distribution into a conditionally Gaussian one by introducing an auxiliary field. The HST is sometimes described as a "trick"—a clever manipulation that appears from nowhere. This description is misleading. The HST is a latent-variable augmentation: an exact identity that introduces a new random variable so that the conditional distribution of the original variable becomes Gaussian.

4.2 The Gaussian Integral Identity

The simplest version of the HST is a one-dimensional Gaussian integral identity. Let O be a real number and $g > 0$ a positive coupling constant. Then:

$$\exp\left(\frac{g}{2} O^2\right) = \frac{1}{\sqrt{2\pi g}} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2}{2g} + \sigma O\right) d\sigma. \tag{4.1}$$

Let us derive this from scratch, so there is no mystery about where it comes from. The right-hand side is a Gaussian integral in σ. The exponent $-\frac{\sigma^2}{2g} + \sigma O$ is a quadratic in σ with precision $1/g$ and source O. By the completing-the-square formula (Section 2.3):

$$-\frac{\sigma^2}{2g} + \sigma O = -\frac{1}{2g}(\sigma - gO)^2 + \frac{g}{2}O^2.$$

The integral over σ of $\exp\left(-\frac{1}{2g}(\sigma - gO)^2\right)$ gives $\sqrt{2\pi g}$ (a Gaussian integral with variance g). Therefore:

$$\frac{1}{\sqrt{2\pi g}} \int \exp\left(-\frac{\sigma^2}{2g} + \sigma O\right) d\sigma = \exp\left(\frac{g}{2}O^2\right) \cdot \frac{\sqrt{2\pi g}}{\sqrt{2\pi g}} = \exp\left(\frac{g}{2}O^2\right).$$

That is all. The identity (4.1) is a Gaussian integral evaluated by completing the square. It is exact—no approximation is involved.

4.3 Why This Identity Matters

Now suppose we have a non-Gaussian distribution with a quartic interaction, and we can write the quartic term in the form $\frac{g}{2}O(\phi)^2$ for some observable $O(\phi)$. Then (4.1) allows us to write

$$e^{-S(\phi)} = e^{-S_0(\phi)} \cdot e^{\frac{g}{2}O(\phi)^2} = e^{-S_0(\phi)} \cdot \frac{1}{\sqrt{2\pi g}} \int \exp\left(-\frac{\sigma^2}{2g} + \sigma\, O(\phi)\right) \mathrm{d}\sigma,$$

where $S_0(\phi)$ contains the quadratic part of the action. In the joint distribution of (ϕ, σ):

$$p(\phi, \sigma) \propto \exp\left(-S_0(\phi) - \frac{\sigma^2}{2g} + \sigma\, O(\phi)\right). \tag{4.2}$$

The original distribution is recovered by marginalization:

$$p(\phi) = \int p(\phi, \sigma)\, \mathrm{d}\sigma.$$

Now consider the conditional distribution of ϕ given σ. If $S_0(\phi)$ is quadratic in ϕ and $O(\phi)$ is linear in ϕ, then the exponent in (4.2) is quadratic in ϕ for any fixed σ, and therefore $p(\phi \mid \sigma)$ is Gaussian. The non-Gaussianity of the original distribution has been absorbed into the coupling between ϕ and the latent variable σ.

This is the essential content of the Hubbard–Stratonovich transformation: a non-Gaussian distribution over ϕ has been expressed as a mixture of Gaussian distributions, one for each value of the latent variable σ.

4.4 Sign Conditions and the Direction of Decoupling

The identity (4.1) requires $g > 0$—the coefficient of O^2 must be positive (an attractive interaction in physics language). If instead we have a negative coefficient $-\frac{g}{2}O^2$ with $g > 0$ (a repulsive interaction), the identity takes the form

$$\exp\left(-\frac{g}{2}O^2\right) = \frac{1}{\sqrt{2\pi g}} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2}{2g} + \mathrm{i}\sigma O\right) \mathrm{d}\sigma, \tag{4.3}$$

where the coupling σO now carries a factor of i. The derivation is the same completing-the-square argument, applied with iO in place of O. The integral is still convergent and exact, but the coupling is imaginary. For bosonic (real) fields in a Euclidean setting, this imaginary coupling generally does not spoil the computation, but it does mean that the latent variable σ cannot be straightforwardly interpreted as a real random variable with a positive measure. We will note such sign issues as they arise.

An important point: the "channel" in which we perform the HST is a choice. A quartic interaction like $\lambda(\phi^2)^2$ can be decoupled by introducing a latent field coupled to ϕ^2 (a density or mass channel), or—with more work—by introducing latent fields coupled to pair operators (a pairing channel), or in other ways. Different channels lead to different latent-variable representations of the same original distribution. They are all exact, but they produce different effective actions for the latent variable, and some may be more amenable to approximation than others.

This freedom is analogous to the choice of latent-variable parameterization in a mixture model or factor model: the same marginal can arise from many different joint distributions.

Worked example: HST for a repulsive interaction

Consider the density $p(\phi) \propto \exp(-\frac{1}{2}m^2\phi^2 - \frac{g}{2}\phi^4)$ with $g > 0$. The quartic term $-\frac{g}{2}\phi^4$ has a negative coefficient: it is a repulsive interaction in the sense that it penalizes large ϕ. We cannot use the identity (4.1) directly because it requires a positive coefficient. Instead, we use (4.3):

$$e^{-\frac{g}{2}\phi^4} = e^{-\frac{g}{2}(\phi^2)^2} = \frac{1}{\sqrt{2\pi g}} \int \exp\left(-\frac{\sigma^2}{2g} + \mathrm{i}\sigma\phi^2\right) \mathrm{d}\sigma.$$

The joint density becomes

$$p(\phi, \sigma) \propto \exp\left(-\frac{1}{2}(m^2 - \mathrm{i}\sigma)\phi^2 - \frac{\sigma^2}{2g}\right).$$

For fixed σ, the ϕ-integral is Gaussian with precision $m^2 - \mathrm{i}\sigma$, provided $\mathrm{Re}(m^2 - \mathrm{i}\sigma) = m^2 > 0$ (which holds for real σ). Integrating out ϕ:

$$S_{\text{eff}}(\sigma) = \frac{\sigma^2}{2g} + \frac{1}{2}\log(m^2 - \mathrm{i}\sigma) + \text{const.}$$

The saddle-point equation $\frac{\partial S_{\text{eff}}}{\partial \sigma} = 0$ gives $\frac{\sigma}{g} = \frac{\mathrm{i}}{2(m^2 - \mathrm{i}\sigma)}$, or $\sigma(m^2 - \mathrm{i}\sigma) = \frac{\mathrm{i}g}{2}$. Expanding: $m^2\sigma - \mathrm{i}\sigma^2 = \frac{\mathrm{i}g}{2}$. Setting $\sigma = \mathrm{i}s$ (rotating the integration contour to the imaginary axis, so that s is real):

$$\mathrm{i}m^2 s + s^2 = \frac{\mathrm{i}g}{2}, \quad \text{i.e.,} \quad s^2 + \mathrm{i}m^2 s - \frac{\mathrm{i}g}{2} = 0.$$

For small g, the solution is $s \approx g/(2m^2)$ (real and positive), giving $\sigma = \mathrm{i}s = \mathrm{i}g/(2m^2)$ and a saddle-dressed mass $m^2 - \mathrm{i}\sigma = m^2 + g/(2m^2)$. The quartic interaction has increased the effective mass, as expected for a repulsive coupling.

The imaginary coupling $\mathrm{i}\sigma$ in the HST is mathematically harmless—it simply means that the saddle lies on the imaginary axis in the σ-plane, and the steepest-descent contour must be deformed accordingly. But it does mean that σ cannot be interpreted as an ordinary random variable with a positive density. We will encounter this again in the electron gas (Chapter 29), where the Coulomb repulsion leads to the same imaginary-coupling structure.

Worked example: comparing two HST channels

The choice of HST channel—which composite observable $O(\phi)$ the auxiliary field σ couples to—is not unique. Different choices are all exact but lead to different effective actions with different saddle-point approximations.

Consider N real fields ϕ_a, $a = 1, \ldots, N$, with a quartic interaction that can be written in two ways:

$$\frac{\lambda}{4N}\left(\sum_a \phi_a^2\right)^2 = \frac{\lambda}{4N}\sum_{a,b} \phi_a^2 \phi_b^2.$$

Channel 1 (density): Introduce a scalar σ coupled to $O_1 = \frac{1}{\sqrt{N}}\sum_a \phi_a^2$. The joint action has ϕ entering quadratically for fixed σ, and integrating out ϕ gives $S_{\text{eff}}^{(1)}[\sigma] = \frac{N}{4\lambda}\sigma^2 + \frac{N}{2}\,\mathrm{Tr}\log(-\nabla^2 + m_0^2 + \sigma)$. This is the channel used throughout most of this book.

Channel 2 (exchange): Alternatively, introduce a matrix M_{ab} coupled to the tensor $O_2^{ab} = \phi_a\phi_b$. Then each ϕ_a sees a different potential from M, and the effective action becomes $S_{\text{eff}}^{(2)}[M] =$

$\frac{N}{4\lambda}\operatorname{Tr} M^2 + \frac{1}{2}\operatorname{Tr}\log(-\nabla^2\delta_{ab} + m_0^2\delta_{ab} + M_{ab})$, where the log-determinant now involves an $N \times N$ matrix kernel.

Both channels give the same exact result after performing the σ or M integral, but the saddle-point approximation is different in each. Channel 1 has a single scalar saddle $\bar{\sigma}$ and captures all N components symmetrically, making it natural for the symmetric (disordered) phase. Channel 2 has a matrix saddle $\bar{M}_{ab}$ that can break the $O(N)$ symmetry explicitly, making it more flexible but harder to solve. In the symmetric phase, both saddles agree: $\bar{M}_{ab} = \bar{\sigma}\,\delta_{ab}$. In the broken phase, Channel 2 can find anisotropic saddles that Channel 1 misses.

The lesson is practical: when setting up an HST, choose a channel whose saddle is likely to capture the dominant physics of the problem. Symmetry considerations often guide the choice: if the phase of interest preserves the symmetry, the simplest (most symmetric) channel usually suffices. If symmetry breaking is expected, a channel that allows the saddle to break the symmetry is preferable.

4.5 Multivariate and Field-Theoretic Generalization

The one-dimensional identity (4.1) generalizes to many dimensions. For a vector $O \in \mathbb{R}^m$ and a positive-definite $m \times m$ coupling matrix G:

$$\exp\left(\frac{1}{2}O^\top G O\right) = \frac{1}{(2\pi)^{m/2}(\det G)^{1/2}} \int \exp\left(-\frac{1}{2}\sigma^\top G^{-1}\sigma + \sigma^\top O\right) \mathrm{d}\sigma. \tag{4.4}$$

In the functional-integral (field theory) notation, O becomes a function $O(x)$, the coupling matrix G becomes a kernel $G(x, y)$, the latent vector σ becomes a latent field $\sigma(x)$, and the integral becomes a functional integral $\int \mathcal{D}\sigma$. Every step is algebraically identical to the finite-dimensional case.

4.6 What the Latent Variable Represents

In the Hubbard–Stratonovich framework, the latent field σ is coupled to the composite observable $O(\phi)$. Its role is to mediate the interaction: instead of having a direct quartic coupling O^2, we have a linear coupling σO between σ and ϕ, and a quadratic "prior" $\frac{\sigma^2}{2g}$ on σ. The interaction is now indirect—ϕ influences σ through the coupling, and σ influences ϕ back.

The bird/flock analogy

A vivid analogy illuminates the entire HST framework. Consider a flock of starlings—hundreds of individual birds, each adjusting its flight in response to its neighbors, producing a coherent, swirling pattern.

The microscopic field ϕ is the individual bird: its position, velocity, and orientation. Each bird interacts with its neighbors through local rules ("fly at the same speed, in the same direction, without colliding"). These local interactions are the quartic (non-Gaussian) terms in the action.

The collective field σ is the flock pattern: the large-scale shape, density, and flow direction of the flock. No single bird *is* the flock, but the flock pattern emerges from the collective behavior of all the birds.

The HST is the mathematical device that introduces the flock pattern as a new variable. Once the flock pattern σ is specified, each bird's behavior becomes conditionally simple (Gaussian): it fluctuates around the local flock direction with some variance. The complexity—the non-Gaussianity—has been moved from the individual birds into the coupling between birds and flock.

An even more apt analogy: **people and society**. Each person ϕ_i is a microscopic degree of freedom, interacting with neighbors through social bonds (the quartic coupling). **Society** σ—the collective norms, institutions, culture—is the latent field. Given societal norms (σ fixed), each person's behavior is relatively predictable ($p(\phi|\sigma)$ is Gaussian). But the norms themselves are not imposed from outside; they emerge self-consistently from the aggregate behavior of all individuals (the gap equation: $\bar{\sigma}$ is determined by $\langle O(\phi)\rangle_{\bar{\sigma}}$).

The saddle $\bar{\sigma}$ is the self-consistent social equilibrium: the norms that, when adopted, produce behavior that sustains those norms.

The **gap** Δ is the cost a person must pay to break free from society. In a tightly conformist society (large Δ), deviating from the norm is expensive: social ostracism, economic penalty, loss of cooperation. The individual's "excitation energy"—the energy needed to create a person whose behavior differs from the collective pattern—is at least Δ. Below this threshold, no individual excitation is possible; everyone remains locked into the collective mode. This is exactly the quasiparticle gap in superconductivity: the energy Δ_0 needed to break a Cooper pair and create an independent electron excitation.

In a society with a small gap (weak conformity pressure), individuals can deviate cheaply, and the collective pattern is fragile—fluctuations are large, correlations are long-ranged, and the society is near a "phase transition" to a different collective state. In a society with a large gap (strong conformity), individuals are tightly bound to the collective mode, fluctuations are suppressed, and the society is rigid.

A **Goldstone mode** is a direction in which the *entire* society can "rotate" without cost: changing the language everyone speaks, or the side of the road everyone drives on. The choice is arbitrary, but once made, it is stable—the saddle has broken a symmetry. No individual pays a cost for the rotation, because everyone rotates together. But if one person alone tries to drive on the other side of the road, the cost is enormous—this is the gap, not the Goldstone mode. The gap measures the cost of *individual* deviation; the Goldstone mode is a *collective* rotation that costs nothing.

At the saddle point of the σ integral (discussed in Part IV), the latent variable takes the value $\bar{\sigma} = g\,\langle O(\phi)\rangle_{\bar{\sigma}}$, which is the expected value of the observable O, rescaled by the coupling g. In physical contexts, $\bar{\sigma}$ becomes an order parameter: a magnetization, a pairing gap, a density background, or a condensate, depending on what O represents.

This is not merely a mathematical convenience. The introduction of a latent variable coupled to a physically meaningful observable—the density, the pair operator, the magnetization—gives us a new field whose fluctuations describe *collective* behavior. The precision (inverse covariance) of this collective field, computed by taking the second derivative of the effective action, is the polarization or susceptibility of the observable O. We will develop this in Part V.

Chapter 5

Effective Interactions from Latent-Variable Marginalization

5.1 Gaussian–Gaussian Latent Models

To see how latent-variable marginalization generates effective interactions, consider a joint Gaussian distribution over two groups of variables, which we call "visible" v and "hidden" h:

$$p(v, h) \propto \exp\left(-\frac{1}{2}v^\top Av - \frac{1}{2}h^\top Bh + v^\top Wh\right), \tag{5.1}$$

where A and B are positive-definite matrices and W is a coupling matrix. This is a Gaussian with a block precision structure.

What is the marginal distribution of v? We integrate out h. The terms involving h are

$$-\frac{1}{2}h^\top Bh + h^\top W^\top v,$$

which is a Gaussian in h with precision B and source $W^\top v$. By completing the square:

$$\int \exp\left(-\frac{1}{2}h^\top Bh + h^\top W^\top v\right) \mathrm{d}h \propto \exp\left(\frac{1}{2}v^\top WB^{-1}W^\top v\right).$$

Combining with the v-dependent terms:

$$p(v) \propto \exp\left(-\frac{1}{2}v^\top \left(A - WB^{-1}W^\top\right) v\right). \tag{5.2}$$

The marginal precision of v is $A - WB^{-1}W^\top$, which is the Schur complement of B in the joint precision matrix, exactly as in Section 2.5. The effect of the hidden variables is to modify the precision of the visible variables by the term $-WB^{-1}W^\top$.

Notice the sign: the marginal precision is *reduced* relative to the conditional precision A. Integrating out the hidden variables has generated an effective attraction (a reduction in cost for certain configurations of v) that was not present in the original visible-variable precision. This is the mechanism by which latent variables generate effective interactions.

5.2 Connection to Restricted Boltzmann Machines

In machine learning, a closely related structure arises in the restricted Boltzmann machine (RBM). An RBM with binary hidden units has the joint energy function

$$E(v, h) = -v^\top W h - b^\top v - c^\top h,$$

where $h_i \in \{-1, +1\}$. The marginal distribution of v is obtained by summing over h:

$$p(v) \propto \sum_h \exp\Big(v^\top W h + b^\top v + c^\top h\Big) .$$

Since the h_i are conditionally independent given v (this is the "restricted" structure), the sum factorizes:

$$\sum_h \exp\Big(v^\top W h + c^\top h\Big) = \prod_j \sum_{h_j = \pm 1} \exp\Big(h_j (w_j^\top v + c_j)\Big) = \prod_j 2 \cosh\Big(w_j^\top v + c_j\Big) ,$$

where w_j is the j-th column of W. Taking the logarithm, the effective action for v is

$$S_{\text{eff}}(v) = -b^\top v - \sum_j \log\cosh\Big(w_j^\top v + c_j\Big) + \text{const.} \tag{5.3}$$

The $\log\cosh$ terms are nonlinear (non-quadratic) functions of v, so the marginal distribution of v is non-Gaussian. The latent marginalization has generated effective interactions among the visible variables.

Compare this to the Gaussian–Gaussian case (5.2), where the marginal is still Gaussian. The difference is that binary hidden units produce $\log\cosh$ nonlinearities instead of the quadratic $\frac{1}{2} v^\top W B^{-1} W^\top v$ that arises from Gaussian hidden units. The algebraic mechanism—marginalization of hidden variables generates effective interactions—is the same.

> Latent marginalization generates effective interactions. Whether the latent variables are continuous Gaussian, binary, Poisson, or anything else, integrating them out produces terms in the effective action that couple the visible variables in new ways. The precise form of the effective interaction depends on the latent distribution.

5.3 Effective Action as Negative Log Marginal Density

The general setting is a joint distribution $p(\phi, \sigma) \propto e^{-S(\phi, \sigma)}$, where ϕ are the variables of primary interest and σ are latent or auxiliary variables. The marginal distribution of ϕ is

$$p(\phi) = \int p(\phi, \sigma)\, \mathrm{d}\sigma \propto \exp(-S_{\text{eff}}(\phi)) ,$$

where the **effective action** is

$$S_{\text{eff}}(\phi) = -\log \int e^{-S(\phi, \sigma)}\, \mathrm{d}\sigma + \text{const.} \tag{5.4}$$

This is simply the negative log marginal density. It equals the negative logarithm of the integral over latent variables.

In statistics, $\int e^{-S(\phi,\sigma)}\,\mathrm{d}\sigma$ is the marginal likelihood (with ϕ playing the role of data or parameters, depending on context). In physics, $S_{\text{eff}}(\phi)$ is the **effective action obtained by integrating out** σ. Only after stating the mathematical definition do we introduce the physics term.

We will use the same structure in the other direction: when the HST introduces a latent field σ and we integrate out the original field ϕ, we obtain an effective action for σ:

$$S_{\text{eff}}(\sigma) = -\log \int e^{-S(\phi,\sigma)}\,\mathrm{d}\phi + \text{const.}$$

The effective action for the latent field σ is the object we will study in most of the rest of the book. Its saddle points, Hessians, and higher derivatives encode the physics of gaps, collective modes, and phase transitions.

5.4 The Law of Total Covariance

A useful identity from probability theory connects the covariance structures before and after marginalization. For a joint distribution $p(\phi, \sigma)$:

$$\text{Cov}(\phi) = \mathbb{E}_\sigma[\text{Cov}(\phi \mid \sigma)] + \text{Cov}_\sigma(\mathbb{E}[\phi \mid \sigma])\,. \tag{5.5}$$

This is the **law of total covariance** (also called the variance decomposition or Eve's law). The first term is the average conditional covariance—the fluctuation of ϕ that remains even when σ is known. The second term is the covariance of the conditional mean—the variability of ϕ that arises because σ itself is random.

In the Gaussian–Gaussian model (5.1), the conditional covariance $\text{Cov}(\phi \mid \sigma) = A^{-1}$ does not depend on σ (a special property of Gaussians), so the first term is simply A^{-1}. The conditional mean $\mathbb{E}[\phi \mid \sigma] = -A^{-1}W\sigma$ is linear in σ, so the second term is $A^{-1}W\ \text{Cov}(\sigma)\,W^\top A^{-1}$. Adding them gives the full marginal covariance of ϕ, which must equal $(A - WB^{-1}W^\top)^{-1}$ by (5.2). One can verify this using the matrix identity for Schur complements—and the agreement is a consistency check on our calculations.

What was exact, what was approximate, and what was conventional?

Exact. The Gaussian integral identity (4.1) and its multivariate generalization (4.4). The marginal distribution (5.2) from the Gaussian–Gaussian model. The effective-action definition (5.4). The law of total covariance (5.5). Nothing in this chapter was approximate.

Conventional. The term "Hubbard–Stratonovich transformation" for the identity (4.1). The term "effective action" for the negative log marginal density (5.4). The description of the HST as a "trick" in some textbooks—we have argued that it is better understood as latent-variable augmentation.

Chapter 6

The Schwinger Effective Action $\Gamma[\phi_{\text{cl}}]$

We have defined two kinds of "effective action" so far: the negative log marginal density $S_{\text{eff}}[\sigma]$ obtained by integrating out microscopic fields (Chapter 5), and the generating functional $W[J] = \log Z[J]$ of connected correlators (Chapter 1). There is a third, due to Schwinger, that connects the two and provides the natural home for the one-loop $\operatorname{Tr}\log$. It is the Legendre transform of $W[J]$, and its Hessian is the *precision* (inverse propagator)—the perfect dual of the covariance.

6.1 The Legendre Transform

Recall that $W[J]$ generates cumulants: $\phi_{\text{cl}}(x) \equiv \frac{\delta W}{\delta J(x)} = \langle \phi(x) \rangle_J$ is the mean field in the presence of the source. Define the **Schwinger effective action** (or 1PI effective action) as the Legendre transform:

$$\boxed{\Gamma[\phi_{\text{cl}}] = \sup_J \{J \cdot \phi_{\text{cl}} - W[J]\} = J_* \cdot \phi_{\text{cl}} - W[J_*],} \tag{6.1}$$

where $J_*[\phi_{\text{cl}}]$ is the source that produces the mean field ϕ_{cl}: it satisfies $\frac{\delta W}{\delta J}\big|_{J_*} = \phi_{\text{cl}}$. This is the same Legendre transform that converts entropy to free energy in thermodynamics, or the log-partition function to the rate function in large deviations theory.

6.2 The Equation of Motion

Differentiating (6.1) with respect to ϕ_{cl}:

$$\frac{\delta \Gamma}{\delta \phi_{\text{cl}}(x)} = J_*(x). \tag{6.2}$$

At $J = 0$ (no external source), the physical vacuum satisfies $\frac{\delta \Gamma}{\delta \phi_{\text{cl}}} = 0$: the Schwinger effective action is stationary at the physical mean field. This is the quantum equation of motion, generalizing the classical Euler–Lagrange equation.

6.3 The Hessian of Γ Is the Precision

The crucial property: the second derivative of Γ is the *inverse* of the second derivative of W. To see this, differentiate $\phi_{\text{cl}}(x) = \frac{\delta W}{\delta J(x)}$ with respect to $\phi_{\text{cl}}(y)$, using the chain rule through J:

$$\delta(x-y) = \int \mathrm{d}z \, \frac{\delta^2 W}{\delta J(x)\, \delta J(z)} \cdot \frac{\delta J_*(z)}{\delta \phi_{\text{cl}}(y)}.$$

But $\frac{\delta^2 W}{\delta J(x)\delta J(z)} = G(x,z)$ (the connected two-point function, i.e., the **covariance**), and $\frac{\delta J_*}{\delta\phi_{\text{cl}}} = \frac{\delta^2\Gamma}{\delta\phi_{\text{cl}}\delta\phi_{\text{cl}}}$ (from (6.2)). Therefore:

$$\boxed{\frac{\delta^2\Gamma}{\delta\phi_{\text{cl}}(x)\,\delta\phi_{\text{cl}}(y)} = G^{-1}(x,y) = (\text{inverse propagator} = \text{precision}).} \tag{6.3}$$

This is the covariance–precision duality of Chapter 2, now elevated to the full quantum theory:

	Second derivative	**Interpretation**
$W[J]$	$\delta^2 W/\delta J^2 = G$	Covariance (propagator)
$\Gamma[\phi_{\text{cl}}]$	$\delta^2\Gamma/\delta\phi_{\text{cl}}^2 = G^{-1}$	Precision (inverse propagator)

W and Γ contain the same information, but W encodes it as covariance (how fields fluctuate together) while Γ encodes it as precision (how costly those fluctuations are).

6.4 The One-Loop Approximation: Where $\operatorname{Tr}\log$ Appears

The Schwinger effective action has a systematic expansion. At tree level (saddle-point), $\Gamma[\phi_{\text{cl}}] \approx S[\phi_{\text{cl}}]$—the classical action evaluated at the mean field. The first quantum correction is the one-loop determinant:

$$\boxed{\Gamma[\phi_{\text{cl}}] = S[\phi_{\text{cl}}] + \frac{1}{2}\operatorname{Tr}\log S''[\phi_{\text{cl}}] + (\text{two-loop and higher}),} \tag{6.4}$$

where $S''[\phi_{\text{cl}}] = \frac{\delta^2 S}{\delta\phi^2}\big|_{\phi_{\text{cl}}}$ is the Hessian of the classical action at ϕ_{cl}.

This is *the* canonical definition of the one-loop effective action and of the $\operatorname{Tr}\log$ in field theory. It is the Laplace approximation (Chapter 8) applied to the path integral $Z[J] = \int \mathcal{D}\phi\, e^{-S[\phi]+J\phi}$, expanded around the saddle ϕ_{cl}.

6.5 Comparison: Schwinger Γ vs. HST S_{eff}

	Schwinger $\Gamma[\phi_{\text{cl}}]$	**HST** $S_{\text{eff}}[\sigma]$
Variable	Mean field $\phi_{\text{cl}} = \langle\phi\rangle_J$	Auxiliary field σ
Defined by	Legendre transform of $W[J]$	Integrating out ϕ from $p(\phi,\sigma)$
Hessian gives	Precision G^{-1} of ϕ	Precision D^{-1} of σ (collective)
One-loop	$S[\phi_{\text{cl}}] + \frac{1}{2}\operatorname{Tr}\log S''[\phi_{\text{cl}}]$	$S_0[\sigma] + \frac{1}{2}\operatorname{Tr}\log M(\sigma)$
Exact for	No specific case	ϕ entering quadratically
Physical content	Full quantum effective potential	Collective-field effective action

The two effective actions are *different objects* that play complementary roles. $\Gamma[\phi_{\text{cl}}]$ is the effective action for the original field ϕ—its Hessian gives the precision (inverse propagator) of ϕ. $S_{\text{eff}}[\sigma]$ is the effective action for the auxiliary field σ—its Hessian gives the precision of σ (the collective propagator). Both involve $\operatorname{Tr}\log$, but the operator inside the $\operatorname{Tr}\log$ is different: $S''[\phi_{\text{cl}}]$ in Schwinger's case, $M(\sigma)$ in the HST case.

6.6 Connecting the Auxiliary-Field Propagator to the Fundamental-Field Propagator

A question that frequently confuses newcomers: if σ is an auxiliary variable introduced by hand, how does its propagator $D = K^{-1}$ relate to the propagator G of the fundamental field ϕ?

The answer involves the law of total covariance (Chapter 5) and the structure of the HST coupling.

The exact relation

In the joint distribution $p(\phi, \sigma)$, the marginal covariance of ϕ is (by the law of total covariance):

$$\mathrm{Cov}(\phi) = \mathbb{E}_\sigma[\mathrm{Cov}(\phi|\sigma)] + \mathrm{Cov}_\sigma(\mathbb{E}[\phi|\sigma]).$$

The first term is the average conditional covariance: the fluctuation of ϕ at fixed σ. For the HST, $p(\phi|\sigma)$ is Gaussian with precision $M(\sigma)$, so $\mathrm{Cov}(\phi|\sigma) = M(\sigma)^{-1} = G_\sigma$.

The second term is the covariance of the conditional mean. If $O(\phi)$ is the observable coupled to σ and $\mathbb{E}[\phi|\sigma]$ depends on σ through G_σ, then this term involves D (the covariance of σ).

At the RPA level (Gaussian approximation for σ around $\bar{\sigma}$), the total covariance of ϕ becomes

$$G_{ij} = [G_{\bar{\sigma}}]_{ij} + \sum_{xy} [G_{\bar{\sigma}}\Gamma]_{ix}\, D(x, y)\, [\Gamma G_{\bar{\sigma}}]_{yj}. \tag{6.5}$$

The first term is the saddle-dressed propagator (the covariance at fixed $\bar{\sigma}$). The second term is the correction from σ-fluctuations: two "legs" of $G_{\bar{\sigma}}\Gamma$ coupling ϕ to σ, connected by the collective propagator D.

This is a Dyson-like equation, but for the full self-energy including collective-field fluctuations:

$$G^{-1} = G_{\bar{\sigma}}^{-1} + \Sigma_{\mathrm{fluc}}, \qquad \Sigma_{\mathrm{fluc}} \sim \Gamma\, D\, \Gamma.$$

The fluctuation self-energy Σ_{fluc} is the contribution from virtual exchange of collective-mode quanta (σ-propagator insertions), dressed by the coupling vertices Γ.

The physical picture

The saddle-dressed propagator $G_{\bar{\sigma}}$ includes the effect of the *mean* collective field (the average societal norm, in the person/society analogy). The correction $\Gamma D \Gamma$ includes the effect of *fluctuations* of the collective field (fluctuations of societal norms around their equilibrium). The full propagator G includes both.

In many systems, the dominant effect is the saddle: $G \approx G_{\bar{\sigma}}$, with the fluctuation correction being a $1/N$ or $1/V$ effect. Near a phase transition, however, D diverges (because $K \to 0$), and the fluctuation correction becomes large. This is why mean-field theory breaks down near critical points: the collective-field propagator D becomes long-ranged, and its contribution to G through the $\Gamma D \Gamma$ insertion dominates.

Diagrammatic interpretation

The relation $G = G_{\bar{\sigma}} + G_{\bar{\sigma}}\Gamma D \Gamma G_{\bar{\sigma}} + \cdots$ has a clear diagrammatic interpretation. Each D insertion is a "bubble chain" (since $D = (D_0^{-1} + \Pi)^{-1}$ is itself an infinite series of bubbles). The full ϕ-propagator includes the bare propagator, the saddle correction (from $\bar{\sigma}$), and an infinite series of collective-mode exchange diagrams. This is exactly the resummation that the RPA performs: it replaces the direct ϕ-ϕ interaction by the exchange of dressed collective quanta.

In particle physics language: the bare interaction V between two electrons is the Coulomb potential. The RPA replaces this by the screened interaction D, which includes the polarization of the medium. The electron propagator G includes both the Hartree self-energy (from $\bar{\sigma}$) and the exchange of plasmons and screened Coulomb quanta (from D). The auxiliary-field propagator D mediates the effective interaction between the fundamental particles.

In the bird/flock analogy: $\Gamma[\phi_{\text{cl}}]$ is the effective action governing the "average bird" ϕ_{cl}—its Hessian tells us how costly it is to change the average bird's behavior. $S_{\text{eff}}[\sigma]$ is the effective action governing the "flock pattern" σ—its Hessian tells us how costly it is to change the collective pattern. The two are related but distinct: changing the average bird changes the flock, and changing the flock changes each bird, but the costs of these two operations are different.

Emergent excitations are not fundamental particles

There is a deeper asymmetry between Γ and S_{eff} that goes beyond the comparison table above. The Schwinger effective action $\Gamma[\phi_{\text{cl}}]$ is defined for the *fundamental* field ϕ—the field that appears in the microscopic Lagrangian and that has its own path integral weight $e^{-S[\phi]}$. The HST effective action $S_{\text{eff}}[\sigma]$ is defined for an *auxiliary* field σ that was introduced by hand through the Hubbard–Stratonovich identity. The field σ does not appear in the original Lagrangian; it has no independent microscopic dynamics. Its entire effective action—including its "kinetic energy" (gradient terms) and "mass" (Hessian eigenvalue)—is *generated* by integrating out the fundamental field ϕ.

This means that $S_{\text{eff}}[\sigma]$ generally cannot be written as a Schwinger effective action for σ. There is no underlying microscopic action $S_0[\sigma]$ for which S_{eff} is the Legendre transform. The collective excitations of σ—phonons, plasmons, magnons, the Higgs mode, the Anderson–Bogoliubov mode—are **emergent**. They propagate, scatter, and decay, and they can be detected experimentally, but they are not fundamental particles in the sense of having their own entry in the microscopic Lagrangian. They exist because of—and only because of—the collective organization of the microscopic degrees of freedom.

In the person/society analogy: "societal norms" are not a person. You cannot write a microscopic Lagrangian for a norm. Norms emerge from the aggregate behavior of people, and they can change, propagate, and interact—but they are a different ontological category from the individuals whose behavior generates them. The Schwinger effective action describes the quantum mechanics of individuals; the HST effective action describes the emergent dynamics of their collective patterns.

6.7 The Self-Energy from Γ

Since $\Gamma''[\phi_{\text{cl}}] = G^{-1}$, and for a free theory $\Gamma''_0 = G_0^{-1} = S''$, the self-energy is

$$\Sigma = \Gamma''[\phi_{\text{cl}}] - S''[\phi_{\text{cl}}] = G^{-1} - G_0^{-1},$$

recovering our definition from Chapter 17. The self-energy is the difference between the full and classical Hessians of Γ—the quantum correction to the precision. Higher derivatives of Γ give the 1PI vertices, which are the irreducible building blocks from which all connected correlators are constructed.

Chapter 7

The HST as a Composite-Operator Source

The Schwinger effective action $\Gamma[\phi_{\text{cl}}]$ and the HST effective action $S_{\text{eff}}[\sigma]$ look like different objects. This chapter shows that they are connected by a simple reformulation: the HST auxiliary field σ is a *source coupled to a composite operator*, and the HST "prior" $\sigma^2/(2g)$ is the *cost of maintaining that source.*

7.1 The Generating Functional for a Composite Operator

Consider a field theory with quadratic action $S_0[\phi]$ and a composite observable $O(\phi)$ (for example, $O(\phi) = \phi(x)^2$). Define the generating functional with a source J coupled to O:

$$W_O[J] = \log \int \mathcal{D}\phi \; \exp\left(-S_0[\phi] + \int J(x)\, O(\phi(x))\, \mathrm{d}x\right). \tag{7.1}$$

This is exactly the Schwinger generating functional from Chapter 1, but with the source coupled to the composite operator O instead of to the fundamental field ϕ. The first derivative $\delta W_O/\delta J(x) = \langle O(\phi(x))\rangle_J$ gives the expectation of O in the presence of the source.

7.2 The HST Effective Action as Source Cost Minus Response Benefit

Now recall the original problem: the full action includes a quartic interaction $-(g/2)\int O(\phi)^2$. The HST introduces an auxiliary field σ to decouple this quartic term:

$$Z = \int \mathcal{D}\sigma \; \exp\left(-\frac{\sigma^2}{2g}\right) \int \mathcal{D}\phi \; \exp\left(-S_0[\phi] + \int \sigma\, O(\phi)\right).$$

The inner ϕ-integral is precisely $e^{W_O[\sigma]}$ — the generating functional (7.1) evaluated at source $J = \sigma$. Therefore:

$$\boxed{S_{\text{eff}}[\sigma] = \frac{\sigma^2}{2g} - W_O[\sigma].} \tag{7.2}$$

The HST effective action is the **source cost** $\sigma^2/(2g)$ minus the **response benefit** $W_O[\sigma]$. The source cost is quadratic: maintaining a source of strength σ costs $\sigma^2/(2g)$. The response benefit is the log-partition function gain from turning on that source. The coupling constant g controls how cheap it is to tilt the distribution: large g means cheap sources and strong effective interactions; small g means expensive sources and weak interactions.

This decomposition reveals the meaning of the HST "prior" $\sigma^2/(2g)$: it is not an arbitrary choice or a mathematical artifact. It is the cost of the source—the price paid for tilting the microscopic distribution toward a particular value of the composite observable O.

7.3 The Gap Equation as an Optimality Condition

The saddle-point condition $\delta S_{\text{eff}}/\delta\sigma = 0$ becomes:

$$\frac{\bar{\sigma}}{g} = \left.\frac{\delta W_O}{\delta J}\right|_{J=\bar{\sigma}} = \langle O(\phi)\rangle_{\bar{\sigma}}. \tag{7.3}$$

The left-hand side is the **marginal cost** of increasing the source strength by $\delta\sigma$. The right-hand side is the **marginal benefit**—the induced expectation of O under the tilted distribution. At the saddle, marginal cost equals marginal benefit.

This is an optimality condition: the source strength $\bar{\sigma}$ is chosen so that the cost of maintaining the source is exactly balanced by the response it induces. In economics language, the saddle is the market-clearing price: the price $\bar{\sigma}$ at which supply (the willingness to pay the source cost $\bar{\sigma}/g$) equals demand (the induced response $\langle O\rangle_{\bar{\sigma}}$).

In the person/society analogy: the "source" σ represents the strength of a social institution (a law, a norm, a cultural practice). Maintaining the institution costs $\sigma^2/(2g)$ (enforcement, education, social pressure). The benefit $W_O[\sigma]$ is the increased social order it produces. The gap equation says: the institution's strength is self-consistently determined at the point where the marginal cost of enforcement equals the marginal benefit of compliance.

7.4 The Hessian and the Collective Susceptibility

The second derivative of S_{eff} gives the collective-field precision:

$$K = \frac{\delta^2 S_{\text{eff}}}{\delta\sigma^2} = \frac{1}{g} - \left.\frac{\delta^2 W_O}{\delta J^2}\right|_{J=\bar{\sigma}} = \frac{1}{g} - \chi_O(\bar{\sigma}),$$

where $\chi_O = \delta^2 W_O/\delta J^2 = \text{Cov}_J(O,O)$ is the susceptibility of O (the covariance of the composite observable under the tilted distribution, evaluated at $J = \bar{\sigma}$). This recovers the Hessian structure $K = D_0^{-1} + \Pi$ from Chapter 15, now with the identification:

$$D_0^{-1} = \frac{1}{g} \quad \text{(inverse source cost)}, \qquad \Pi = -\chi_O \quad \text{(negative susceptibility)}.$$

The collective propagator $D = K^{-1} = (1/g - \chi_O)^{-1}$ diverges when $\chi_O \to 1/g$: when the susceptibility of O is large enough to overcome the source cost, the collective field becomes critical and long-range order develops.

7.5 Connection to Legendre Transforms

If we define the Legendre transform of $W_O[J]$ as

$$\Gamma_O[\sigma_{\text{cl}}] = \sup_J \{J \cdot \sigma_{\text{cl}} - W_O[J]\},$$

where $\sigma_{\text{cl}} = \langle O \rangle_J$ is the mean of O in the tilted distribution, then the HST effective action at the saddle can be written as

$$S_{\text{eff}}[\bar{\sigma}] = \frac{\bar{\sigma}^2}{2g} - W_O[\bar{\sigma}] = \frac{\bar{\sigma}^2}{2g} - \bar{\sigma} \cdot \bar{\sigma}_{\text{cl}} + \Gamma_O[\bar{\sigma}_{\text{cl}}],$$

using $W_O[\bar{\sigma}] = \bar{\sigma} \cdot \bar{\sigma}_{\text{cl}} - \Gamma_O[\bar{\sigma}_{\text{cl}}]$. This shows how the HST effective action, the Schwinger-type Legendre transform Γ_O for the composite operator, and the source cost $\sigma^2/(2g)$ fit together. The HST is the composite-operator generalization of the Schwinger formalism, with an explicit source cost that determines the coupling strength.

Part III

The Origins of Tr log

The expression $\operatorname{Tr} \log M$, where M is a matrix (or operator) that depends on parameters, is one of the most frequently encountered objects in field theory. It appears in the effective action whenever Gaussian variables are integrated out, and also in the Laplace (saddle-point) approximation for non-Gaussian integrals.

This part is devoted to understanding where $\operatorname{Tr} \log M$ comes from, what it means, and how to differentiate it. We will carefully distinguish three distinct origins of the same formula: exact bosonic Gaussian integration, exact fermionic Grassmann integration, and approximate Laplace expansion.

Chapter 8

Exact Bosonic Gaussian Marginalization

8.1 The Setting

Consider a joint distribution of two variables: a "slow" or "collective" variable σ and a "fast" or "microscopic" variable χ. Suppose the joint action has the form

$$S(\chi, \sigma) = S_0(\sigma) + \frac{1}{2}\chi^\top M(\sigma)\chi, \tag{8.1}$$

where $M(\sigma)$ is a positive-definite matrix (or operator) that depends on σ, and the χ-dependence is purely quadratic. The variable χ is, for each fixed σ, a Gaussian random variable.

This is exactly the situation produced by the Hubbard–Stratonovich transformation when we introduce a latent field σ and the original field ϕ enters quadratically. It also arises naturally when a field theory has two types of variables, one of which appears only quadratically (as fermions often do, as we will see in the next chapter).

8.2 Integrating Out the Gaussian Variable

The effective action for σ is obtained by integrating out χ:

$$e^{-S_{\text{eff}}(\sigma)} = \int \mathrm{d}\chi\, \exp\left(-S_0(\sigma) - \frac{1}{2}\chi^\top M(\sigma)\chi\right).$$

Since $S_0(\sigma)$ does not depend on χ, it factors out:

$$e^{-S_{\text{eff}}(\sigma)} = e^{-S_0(\sigma)} \int \mathrm{d}\chi\, \exp\left(-\frac{1}{2}\chi^\top M(\sigma)\chi\right).$$

The χ-integral is a Gaussian integral (equation 2.3):

$$\int \mathrm{d}\chi\, \exp\left(-\frac{1}{2}\chi^\top M(\sigma)\chi\right) = (2\pi)^{n/2} \left(\det M(\sigma)\right)^{-1/2}.$$

Taking $-\log$ of both sides:

$$\boxed{S_{\text{eff}}(\sigma) = S_0(\sigma) + \frac{1}{2}\operatorname{Tr}\log M(\sigma) + \text{const.}} \tag{8.2}$$

Here we used the identity $\log\det M = \operatorname{Tr}\log M$ (which holds because $\det M = \prod_i \lambda_i$ and $\log\prod_i \lambda_i = \sum_i \log\lambda_i = \operatorname{Tr}\log M$ when $\log M$ is defined through the eigendecomposition).

Let us be very precise about the status of this result:

This Tr log is exact. It arises from exact Gaussian integration over χ. No Taylor expansion, no saddle-point approximation, and no truncation have been made. The only assumption is that χ enters the action quadratically.

The constant absorbs the $(2\pi)^{n/2}$ factor and is independent of σ. It can be dropped when we are interested in the σ-dependence of the effective action.

8.3 The Structure of the Effective Action

The effective action (8.2) has two terms. The first, $S_0(\sigma)$, is the "bare" action for σ—whatever σ-dependent terms were present before integrating out χ. The second, $\frac{1}{2}\operatorname{Tr}\log M(\sigma)$, is the contribution from the χ fluctuations, often called the **fluctuation determinant** or **one-loop determinant** (we will explain the "one-loop" terminology later).

The sign is important: the effective action has $+\frac{1}{2}\operatorname{Tr}\log M$, which means that larger values of $\det M$ increase the effective action and therefore suppress the probability weight $e^{-S_{\text{eff}}}$. Intuitively: when $M(\sigma)$ is large (high precision for χ), the Gaussian integral over χ gives a small result (the χ fluctuations are tightly constrained), contributing less to the marginal weight at that value of σ.

Conversely, when $M(\sigma)$ is small (low precision, large covariance), the Gaussian integral gives a large result (the χ fluctuations are broadly distributed), contributing more to the marginal weight. This is the Bayesian "Occam factor"—wider parameter spaces are penalized less by the marginal likelihood.

Chapter 9

Exact Fermionic Gaussian Marginalization

9.1 Why a Separate Chapter?

In physics, many of the variables one wishes to integrate out are fermionic—they describe electrons, quarks, or other spin-$\frac{1}{2}$ particles. Fermionic fields are represented mathematically by **Grassmann variables**: anticommuting algebraic objects that are not ordinary real or complex numbers. This chapter introduces Grassmann variables and shows how their integration produces a $\operatorname{Tr}\log$ with the opposite sign from the bosonic case.

9.2 Grassmann Variables: The Essentials

A Grassmann variable θ is a formal algebraic object satisfying $\theta^2 = 0$. This implies that for two Grassmann variables θ_1 and θ_2: $\theta_1\theta_2 = -\theta_2\theta_1$ (**anticommutativity**). Any function of a single Grassmann variable has only two terms: $f(\theta) = a + b\theta$.

Integration over Grassmann variables is defined algebraically by the **Berezin integral**:

$$\int \mathrm{d}\theta\, 1 = 0, \qquad \int \mathrm{d}\theta\, \theta = 1.$$

These rules may look strange, but they are the unique linear rules that make the integral translation-invariant.

9.3 The Fermionic Gaussian Integral

For n pairs of Grassmann variables $\bar{\psi}_i, \psi_i$ and an $n \times n$ matrix M:

$$\int \mathrm{d}\bar{\psi}_1\, \mathrm{d}\psi_1 \cdots \mathrm{d}\bar{\psi}_n\, \mathrm{d}\psi_n\, \exp\left(-\sum_{i,j} \bar{\psi}_i M_{ij} \psi_j\right) = \det M. \tag{9.1}$$

The crucial difference from the bosonic case: the fermionic integral gives $\det M$ rather than $(\det M)^{-1/2}$.

Let us verify for $n = 1$: a single pair $\bar{\psi}, \psi$ and a scalar m. Then $e^{-m\bar{\psi}\psi} = 1 - m\bar{\psi}\psi$ (because $(\bar{\psi}\psi)^2 = 0$). Therefore $\int \mathrm{d}\bar{\psi}\, \mathrm{d}\psi\, (1 - m\bar{\psi}\psi) = m$, which equals $\det M = m$.

9.4 The Fermionic Effective Action

Integrating out fermions from $S(\psi, \bar{\psi}, \sigma) = \bar{\psi} M(\sigma) \psi + S_0(\sigma)$:

$$\boxed{S_{\text{eff}}(\sigma) = S_0(\sigma) - \operatorname{Tr} \log M(\sigma) + \text{const.}} \tag{9.2}$$

Comparing with the bosonic result (8.2):

$$\text{Bosonic: } S_{\text{eff}} = S_0 + \tfrac{1}{2} \operatorname{Tr} \log M, \qquad \text{Fermionic: } S_{\text{eff}} = S_0 - \operatorname{Tr} \log M.$$

The differences are: (1) **Sign**: bosonic gives $+\operatorname{Tr}\log$, fermionic gives $-\operatorname{Tr}\log$. (2) **Factor of** $\frac{1}{2}$: present for bosonic, absent for fermionic (because a real bosonic Gaussian has $\frac{1}{2}\chi^\top M \chi$ while a fermionic Gaussian has $\bar{\psi} M \psi$ with two independent Grassmann fields).

Both $\operatorname{Tr}\log$ expressions are exact, contingent on the integrated variables entering the action quadratically.

9.5 A Cautionary Note

Grassmann integration is an algebraic operation, not an integral over a probability measure. The "density" $e^{-\bar{\psi} M \psi}$ does not define a probability distribution in the ordinary sense. The fermionic "propagator" M^{-1} is defined by analogy with the bosonic covariance, and it obeys similar algebraic identities (Wick's theorem has a fermionic version with sign factors), but it is not a covariance matrix of an ordinary random variable.

We will use probabilistic language for fermionic quantities where it is algebraically justified, and we will flag explicitly when the analogy breaks down.

Chapter 10

The Laplace Approximation

10.1 When the Exact Route Is Unavailable

The results of the previous two chapters produce Tr log from exact Gaussian integration. But in many problems, the variables we wish to integrate out do not enter the action quadratically. In such cases, we cannot perform the integral exactly. The most widely used approximation is the **Laplace approximation** (also known as the saddle-point approximation), which approximates the integral by expanding the exponent to second order around its maximum.

10.2 The Laplace Approximation in One Dimension

Consider the integral $Z = \int_{-\infty}^{\infty} e^{-Nf(x)}\,\mathrm{d}x$, where $f(x)$ has a unique global minimum at $x = x_*$ with $f'(x_*) = 0$ and $f''(x_*) > 0$. When N is large, the integrand is sharply peaked near x_*, and we approximate f by its second-order Taylor expansion:

$$f(x) \approx f(x_*) + \frac{1}{2}f''(x_*)(x - x_*)^2.$$

Substituting:

$$Z \approx e^{-Nf(x_*)} \int_{-\infty}^{\infty} \exp\left(-\frac{N}{2}f''(x_*)(x - x_*)^2\right)\,\mathrm{d}x = e^{-Nf(x_*)}\sqrt{\frac{2\pi}{Nf''(x_*)}}.$$

Taking $-\log$:

$$-\log Z \approx Nf(x_*) + \frac{1}{2}\log\bigl(Nf''(x_*)\bigr) - \frac{1}{2}\log(2\pi). \tag{10.1}$$

10.3 The Multivariate Laplace Approximation

For a vector $\phi \in \mathbb{R}^n$ and action $S(\phi)$ with a unique minimum at ϕ_* (where $\nabla S(\phi_*) = 0$ and the Hessian $H_* = \nabla^2 S(\phi_*)$ is positive-definite), we claim:

$$Z = \int e^{-S(\phi)}\,\mathrm{d}\phi \approx e^{-S(\phi_*)} \cdot (2\pi)^{n/2}(\det H_*)^{-1/2}. \tag{10.2}$$

Taking $-\log$:

$$-\log Z \approx S(\phi_*) + \frac{1}{2}\operatorname{Tr}\log H_* + \text{const}. \tag{10.3}$$

Let us derive (10.2) step by step. Write $\phi = \phi_* + \eta$ and expand S to second order:

$$S(\phi_* + \eta) = S(\phi_*) + \nabla S(\phi_*)^\top \eta + \frac{1}{2}\eta^\top H_* \eta + O(|\eta|^3).$$

The linear term vanishes because ϕ_* is a critical point: $\nabla S(\phi_*) = 0$ by assumption. Since $\mathrm{d}\phi = \mathrm{d}\eta$ (a translation does not change the measure), the integral becomes

$$Z = \int e^{-S(\phi_*+\eta)}\,\mathrm{d}\eta \approx e^{-S(\phi_*)} \int \exp\left(-\frac{1}{2}\eta^\top H_* \eta\right)\,\mathrm{d}\eta.$$

The remaining integral is exactly the Gaussian integral (2.3) with $A = H_*$:

$$\int \exp\left(-\frac{1}{2}\eta^\top H_* \eta\right)\,\mathrm{d}\eta = (2\pi)^{n/2}(\det H_*)^{-1/2}.$$

Combining gives (10.2). Taking $-\log$ and using $\log\det H_* = \operatorname{Tr}\log H_*$:

$$-\log Z \approx S(\phi_*) + \frac{1}{2}\log\det H_* + \text{const} = S(\phi_*) + \frac{1}{2}\operatorname{Tr}\log H_* + \text{const}.$$

The first term $S(\phi_*)$ is the value of the exponent at the saddle (the MAP value in Bayesian language). The second term $\frac{1}{2}\operatorname{Tr}\log H_*$ is the Laplace correction, arising from the curvature at the saddle.

Here is the $\operatorname{Tr}\log$ again, but now its origin is different from the exact Gaussian case. When applied to the path integral $Z[J] = \int \mathcal{D}\phi\, e^{-S[\phi]+J\phi}$ expanded around the mean field ϕ_{cl}, this Laplace result gives the **one-loop Schwinger effective action**:

$$\Gamma^{(\text{1-loop})}[\phi_{\text{cl}}] = S[\phi_{\text{cl}}] + \frac{1}{2}\operatorname{Tr}\log S''[\phi_{\text{cl}}].$$

This is the canonical definition of the one-loop approximation in quantum field theory: the classical action plus the $\operatorname{Tr}\log$ of its Hessian. The Schwinger effective action $\Gamma[\phi_{\text{cl}}]$ (which we will develop fully in Chapter 6) is defined exactly by a Legendre transform; the expression above is its leading approximation.

Statistics and ML	**Physics**
Mode / MAP point	Saddle / classical background
Inverse Hessian at the mode	Gaussian fluctuation propagator
Laplace correction $\frac{1}{2}\operatorname{Tr}\log H_*$	One-loop determinant
Local Gaussian approximation	One-loop approximation

10.4 The Crucial Distinction

It is essential to distinguish the two origins of $\operatorname{Tr}\log$ that we have now encountered:

Exact Gaussian integration (previous chapters): when the integrated variable enters the action exactly quadratically, the $\operatorname{Tr}\log$ is exact. No approximation has been made.

Laplace approximation (this chapter): when the integrated variable enters the action non-quadratically, the $\operatorname{Tr}\log$ arises from a second-order Taylor expansion and is an approximation. Its accuracy depends on how sharply peaked the integrand is.

Both produce the same algebraic expression, but their status is different. Physics textbooks sometimes blur this distinction, using the phrase "one-loop effective action" for both cases. We will always state which origin is at play.

Chapter 11

Matrix Calculus for $\mathrm{Tr}\log$

The effective action involves $\mathrm{Tr}\log M(\sigma)$, where M depends on a parameter σ. To find saddle points, compute Hessians, and derive gap equations and polarization bubbles, we need to differentiate $\mathrm{Tr}\log M$ with respect to σ. This chapter develops the necessary identities from first principles.

11.1 Derivative of a Matrix Inverse

We begin with the most basic identity. Let $M = M(\sigma)$ be invertible and depend smoothly on σ. Start from $MM^{-1} = I$. Differentiate both sides with respect to σ using the product rule:

$$\frac{dM}{d\sigma}M^{-1} + M\frac{dM^{-1}}{d\sigma} = 0.$$

Solving for $\frac{dM^{-1}}{d\sigma}$:

$$\boxed{\frac{dM^{-1}}{d\sigma} = -M^{-1}\frac{dM}{d\sigma}M^{-1}.} \tag{11.1}$$

This is the matrix analogue of $\frac{d}{dx}\frac{1}{f} = -\frac{f'}{f^2}$. The two factors of M^{-1} (one on each side of $dM/d\sigma$) reflect the non-commutativity of matrix multiplication. In variational notation:

$$\delta(M^{-1}) = -M^{-1}(\delta M)\, M^{-1}. \tag{11.2}$$

This identity will appear repeatedly. It tells us that the derivative of a propagator (covariance) with respect to a parameter that changes the precision is a "sandwich" of two propagators around the precision variation.

11.2 Jacobi's Formula: Derivative of $\mathrm{Tr}\log M$

We derive the identity

$$\boxed{\frac{d}{d\sigma}\,\mathrm{Tr}\log M = \mathrm{Tr}\left(M^{-1}\frac{dM}{d\sigma}\right).} \tag{11.3}$$

Derivation. Consider a family $M(\sigma)$ and write, for small ϵ:

$$M(\sigma+\epsilon) = M(\sigma)\left(I + \epsilon\, M(\sigma)^{-1}\frac{dM}{d\sigma} + O(\epsilon^2)\right).$$

Taking log det of both sides:

$$\operatorname{Tr}\log M(\sigma+\epsilon) = \operatorname{Tr}\log M(\sigma) + \operatorname{Tr}\log\left(I + \epsilon\, M^{-1}\frac{dM}{d\sigma}\right) + O(\epsilon^2).$$

Using $\operatorname{Tr}\log(I+\epsilon X) = \epsilon\ \operatorname{Tr}(X) + O(\epsilon^2)$:

$$\operatorname{Tr}\log M(\sigma+\epsilon) - \operatorname{Tr}\log M(\sigma) = \epsilon\ \operatorname{Tr}\left(M^{-1}\frac{dM}{d\sigma}\right) + O(\epsilon^2).$$

Dividing by ϵ and taking $\epsilon \to 0$ gives (11.3). In variational notation:

$$\delta\ \operatorname{Tr}\log M = \operatorname{Tr}(M^{-1}\delta M). \tag{11.4}$$

This is the identity used to derive gap equations: the saddle-point condition $\frac{\delta S_{\text{eff}}}{\delta\sigma} = 0$ requires differentiating $\operatorname{Tr}\log M(\sigma)$ with respect to σ, and (11.3) tells us that the answer involves the propagator M^{-1}.

11.3 Second Derivative of Tr log

Taking one more derivative gives the Hessian of $\operatorname{Tr}\log M$, which we will identify (in Part V) as the polarization bubble.

Starting from $\frac{d}{d\sigma}\operatorname{Tr}\log M = \operatorname{Tr}(M^{-1}M')$ (primes denote σ-derivatives), differentiate again:

$$\frac{d^2}{d\sigma^2}\operatorname{Tr}\log M = \operatorname{Tr}\left(\frac{d(M^{-1})}{d\sigma}M' + M^{-1}M''\right). \tag{11.5}$$

Using (11.1) for $\frac{d(M^{-1})}{d\sigma}$:

$$= \operatorname{Tr}\big(-M^{-1}M'M^{-1}M' + M^{-1}M''\big) = -\operatorname{Tr}(M^{-1}M'\,M^{-1}M') + \operatorname{Tr}(M^{-1}M''). \tag{11.6}$$

Now, in many applications, $M(\sigma)$ depends **linearly** on σ:

$$M(\sigma) = A_0 + \sigma\Gamma, \tag{11.7}$$

where A_0 and Γ are fixed matrices. In this case, $M' = \Gamma$ and $M'' = 0$. The second term in (11.6) vanishes, leaving

$$\boxed{\frac{d^2}{d\sigma^2}\operatorname{Tr}\log M = -\operatorname{Tr}(G\Gamma\, G\Gamma),} \tag{11.8}$$

where $G = M^{-1}$ is the propagator. This is a product of two propagators and two insertion matrices Γ, traced over all indices. When drawn as a diagram, it consists of two propagator lines running in a loop with insertions Γ at the two junctions—a **bubble**.

We have now derived the bubble from matrix calculus, without any reference to Feynman diagrams or perturbation theory. It is simply the second derivative of $\operatorname{Tr}\log M$ with respect to a parameter on which M depends linearly.

In the functional (field-theory) setting, $\sigma_\alpha \to \sigma(x)$ and the second functional derivative becomes

$$\frac{\delta^2}{\delta\sigma(x)\,\delta\sigma(y)}\operatorname{Tr}\log M = -\operatorname{tr}[G(x,y)\Gamma\, G(y,x)\Gamma]\,, \tag{11.9}$$

where tr denotes a trace over internal indices and $G(x,y)$ is the kernel of M^{-1}. In Fourier space:

$$\Pi(q) = -\int_k \operatorname{tr}[G(k+q)\Gamma\, G(k)\Gamma]\,. \tag{11.10}$$

This is the polarization bubble, derived here as a Hessian.

11.4 The Power-Series Expansion of Tr log

For later reference, we record the expansion of Tr log as a power series. Let $M = A + V$. Then

$$\operatorname{Tr}\log(A+V) = \operatorname{Tr}\log A + \operatorname{Tr}\log(I + A^{-1}V).$$

Using $\log(I+X) = X - \frac{1}{2}X^2 + \frac{1}{3}X^3 - \cdots$ (valid for $\|X\| < 1$):

$$\operatorname{Tr}\log(I + A^{-1}V) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \operatorname{Tr}(A^{-1}V)^n. \tag{11.11}$$

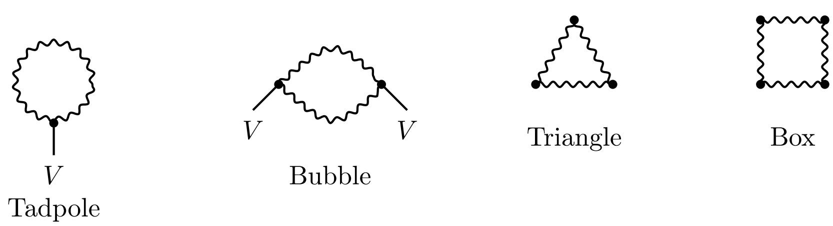
Writing $G_0 = A^{-1}$:

$$\operatorname{Tr}\log(A+V) = \operatorname{Tr}\log A + \operatorname{Tr}(G_0V) - \frac{1}{2}\operatorname{Tr}(G_0VG_0V) + \frac{1}{3}\operatorname{Tr}(G_0VG_0VG_0V) - \cdots \tag{11.12}$$

The terms have a clear algebraic structure: the n-th term is a trace of n alternating propagators and insertions, forming a closed chain.

Order n	**Algebraic expression**	**Physics name**
1	$\operatorname{Tr}(G_0V)$	Tadpole
2	$\frac{1}{2}\operatorname{Tr}(G_0VG_0V)$	Bubble
3	$\frac{1}{3}\operatorname{Tr}(G_0VG_0VG_0V)$	Triangle
4	$\frac{1}{4}\operatorname{Tr}(G_0VG_0VG_0VG_0V)$	Box

Each name refers to the shape of the corresponding diagram: a loop with one, two, three, or four external insertions.



Wavy lines represent propagators G_0; dots represent insertions V. Each diagram is a closed polygon with n sides (propagators) and n corners (insertions). But these are just the terms in a Taylor expansion of $\log(I+X)$, applied to the matrix $X = G_0V$. The algebra is elementary. The diagrams are bookkeeping.

What was exact, what was approximate, and what was conventional?

Exact. Every identity in this chapter—equations (11.1) through (11.10)—is an exact result of finite-dimensional matrix calculus.

Approximate. The power-series expansion (11.11) is exact within its radius of convergence; outside this radius, it becomes a formal or asymptotic series. In continuum field theory, the individual terms may be divergent and require regularization.

Conventional. The names "tadpole," "bubble," "triangle," and "box" for the terms in the Tr log expansion. These are diagrammatic labels for algebraic expressions.

Chapter 12

The Heat Kernel: Evaluating $\mathrm{Tr}\log$ for Differential Operators

The previous chapters showed where $\mathrm{Tr}\log M$ comes from and how to differentiate it. This chapter addresses how to *evaluate* $\mathrm{Tr}\log M$ when M is a differential operator—such as $M = -\nabla^2 + m^2 + V(x)$—acting on functions over a d-dimensional space.

The answer is the **heat kernel expansion**, which expresses $\mathrm{Tr}\log M$ as a series of local geometric invariants. This expansion is the bridge between the abstract $\mathrm{Tr}\log$ and concrete physical quantities (masses, couplings, curvatures).

12.1 The Heat Equation and Its Kernel

Consider $D = -\nabla^2 + V(x)$ on $\mathbb{R}^d$, with $V(x) \geq 0$. The **heat equation** is the diffusion equation $\partial u/\partial t = -Du$, with solution $u(t, x) = (e^{-tD}u_0)(x)$. The integral kernel

$$K(t, x, y) = \langle x|e^{-tD}|y\rangle \tag{12.1}$$

is the **heat kernel**: the temperature at x at time t given a unit heat source at y at time 0.

The name "heat kernel" comes from this physical interpretation. In probability, $K(t, x, y)$ is the transition density of Brownian motion with drift V. In machine learning, $K(t, x, y)$ is a **diffusion kernel**—a positive-definite kernel defining a Gaussian process whose covariance reflects the geometry of the underlying space through D.

12.2 Schwinger's Proper-Time Representation

Julian Schwinger introduced a representation connecting the heat kernel to $\mathrm{Tr}\log$. The key identity is

$$\log\lambda = -\int_0^\infty \frac{dt}{t}\left(e^{-t\lambda} - e^{-t}\right), \qquad \lambda > 0, \tag{12.2}$$

verified by differentiating both sides: $1/\lambda = \int_0^\infty e^{-t\lambda}\,\mathrm{d}t$. Applying to each eigenvalue λ_k of D and summing:

$$\boxed{\mathrm{Tr}\log D = -\int_0^\infty \frac{dt}{t}\left(\mathrm{Tr}\ e^{-tD} - \mathrm{Tr}\ e^{-tD_0}\right),} \tag{12.3}$$

where D_0 is a reference operator providing the subtraction. Schwinger called t the **proper time**: it parameterizes a fictitious diffusion process in which a random walker probes the potential $V(x)$ and the geometry. Short t probes the ultraviolet; long t probes the infrared.

12.3 The Short-Time Expansion: Seeley–DeWitt Coefficients

The central result is the asymptotic expansion of $K(t,x,x)$ as $t \to 0^+$:

$$\boxed{K(t,x,x) = \frac{1}{(4\pi t)^{d/2}} \sum_{n=0}^{\infty} a_n(x)\, t^n.} \tag{12.4}$$

The $a_n(x)$ are the **Seeley–DeWitt coefficients**: local functions of V and its derivatives.

Derivation of a_0. For $V = 0$, the free heat kernel at coincident points is $K_0(t,x,x) = 1/(4\pi t)^{d/2}$, giving $a_0 = 1$.

Derivation of a_1. Using the Duhamel (interaction picture) expansion to first order in V:

$$e^{-t(-\nabla^2+V)} = e^{t\nabla^2} - \int_0^t \mathrm{d}s\, e^{(t-s)\nabla^2} V\, e^{s\nabla^2} + O(V^2).$$

The first-order correction at coincident points gives $\delta K(t,x,x) = -tV(x)K_0(t,x,x) + O(t^2)$, so

$$a_1(x) = V(x). \tag{12.5}$$

A note on sign conventions. The signs require care. For constant $V = m^2$, the exact heat kernel is $K(t,x,x) = (4\pi t)^{-d/2} e^{-m^2 t}$, and the expansion $e^{-m^2 t} = 1 - m^2 t + \frac{m^4 t^2}{2} - \cdots$ has alternating signs. The convention (12.4) defines a_n as the coefficient of t^n in the *asymptotic series* of $K(t,x,x) \cdot (4\pi t)^{d/2}$: so $a_0 = 1$, $a_1 = V$, $a_2 = V^2/2 + \nabla^2 V/6$, with the sign of the exponential absorbed into the definition. When substituted into Schwinger's formula (12.3), the alternating signs of the exponential combine with the t-integration to produce the correct signs in the effective action. The reader should verify consistency by checking the constant-V case: $\frac{1}{2}\operatorname{Tr}\log(-\nabla^2 + m^2) = \frac{\mathrm{Vol}}{2}\int_k \log(k^2+m^2)$, which is the sum of logs of eigenvalues $k^2 + m^2 > 0$, all positive.

The coefficient a_2. At second order, including gradient corrections:

$$a_2(x) = \frac{1}{2}V(x)^2 + \frac{1}{6}\nabla^2 V(x). \tag{12.6}$$

The $V^2/2$ comes from V acting twice; the $\nabla^2 V/6$ from the non-commutativity of $-\nabla^2$ and V. On a curved manifold with Ricci scalar R, an additional term $(\xi - 1/6)R$ appears in a_1, where ξ is the scalar-curvature coupling.

12.4 Heat Kernel on Curved Manifolds: Gravity from Fluctuations

The heat kernel expansion becomes dramatically richer on a curved Riemannian manifold (M, g). Consider the Laplace-type operator

$$D = -\nabla^2 + V(x) + \xi R(x),$$

where $\nabla^2 = g^{-1/2}\partial_\mu(g^{1/2}g^{\mu\nu}\partial_\nu)$ is the Laplace–Beltrami operator, $R(x)$ is the Ricci scalar curvature, and ξ is a coupling constant ($\xi = 0$ for minimal coupling, $\xi = (d-2)/(4(d-1))$ for conformal coupling).

The Seeley–DeWitt expansion on a curved background gives

$$K(t,x,x) = \frac{1}{(4\pi t)^{d/2}} \sum_{n=0}^{\infty} a_n(x)\, t^n, \tag{12.7}$$

where the coefficients now involve curvature invariants:

$a_0(x) = 1$. (Unchanged from flat space.)

$a_1(x) = V + (\xi - \frac{1}{6})R$. The Ricci scalar enters through the commutator $[\nabla_\mu, \nabla_\nu]$ acting on scalars, which produces Christoffel symbols and ultimately R.

$a_2(x) = \frac{1}{2}V^2 + \frac{1}{6}\nabla^2 V + (\xi - \frac{1}{6})RV + \frac{1}{180}(R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} - R_{\mu\nu}R^{\mu\nu}) + \frac{1}{2}(\xi - \frac{1}{6})^2 R^2 + \frac{1}{6}(\frac{1}{5} - \xi)\nabla^2 R$.

Here $R_{\mu\nu\rho\sigma}$ is the Riemann tensor and $R_{\mu\nu}$ is the Ricci tensor. The specific combinations of curvature invariants are universal for a scalar field; they change for fields with spin.

Substituting into Schwinger's formula, the one-loop effective action of a scalar field on a curved background is

$$\frac{1}{2}\operatorname{Tr}\log D = \frac{1}{2(4\pi)^{d/2}} \int \mathrm{d}^d x\, \sqrt{g} \left[\frac{\Lambda^d}{d/2}\, a_0 + \frac{\Lambda^{d-2}}{(d-2)/2}\, a_1 + \cdots + \log\Lambda\, a_{d/2} + \text{finite} \right]. \tag{12.8}$$

The curvature-dependent terms generate gravitational actions:

The $\int \sqrt{g}\, a_0\, \mathrm{d}^d x = \int \sqrt{g}\, \mathrm{d}^d x$ term is the **cosmological constant**: a volume-proportional contribution to the action, independent of curvature.

The $\int \sqrt{g}\, a_1\, \mathrm{d}^d x \supset (\xi - \frac{1}{6}) \int \sqrt{g}\, R\, \mathrm{d}^d x$ term is the **Einstein–Hilbert action**. This is the central result: integrating out a quantum scalar field on a curved background generates the Einstein–Hilbert action $\frac{1}{16\pi G} \int \sqrt{g}\, R\, \mathrm{d}^d x$, where the induced Newton's constant is $G^{-1} \propto (\xi - \frac{1}{6})\Lambda^{d-2}$.

The $\int \sqrt{g}\, a_2\, \mathrm{d}^d x \supset \int \sqrt{g}(c_1 R^2 + c_2 R_{\mu\nu}R^{\mu\nu} + c_3 R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma})\, \mathrm{d}^d x$ terms are **higher-derivative gravity**—quadratic curvature corrections that are important at short distances but suppressed by $1/\Lambda^2$ relative to the Einstein–Hilbert term at long distances.

This is a remarkable result from the perspective of this book's framework: gravity—the Einstein–Hilbert action governing the dynamics of spacetime geometry—*emerges* from the Tr log of matter fields integrated out on a curved background. The mechanism is identical to the one we have seen throughout: integrating out microscopic (matter) degrees of freedom generates an effective action for the collective (geometric) variable, with the Tr log producing both local terms (cosmological constant, Einstein–Hilbert) and non-local terms (through the momentum dependence of the polarization).

Of course, this does not mean gravity is "explained" by quantum matter fluctuations alone—the classical Einstein–Hilbert action exists independently. But it shows that the same mathematical machinery (marginalization → Tr log → effective action) that produces gaps, Goldstone modes, and screening also produces the structure of general relativity.

12.5 Heat Kernel for Gauge-Covariant Laplacians: Yang–Mills from Fluctuations

The heat kernel expansion is even richer when the field carries internal (gauge) indices. Consider a field $\phi^a(x)$ transforming under a gauge group G, coupled to a background gauge field $A^a_\mu(x)$. The gauge-covariant Laplacian is

$$D = -(D_\mu D^\mu) + V(x), \qquad D_\mu = \partial_\mu + \mathrm{i}A_\mu,$$

where $A_\mu = A_\mu^a T^a$ is the gauge connection (a matrix in the Lie algebra of G) and D_μ is the gauge-covariant derivative.

The key new ingredient is the **field strength**

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i}[A_\mu, A_\nu],$$

which enters through the commutator of covariant derivatives: $[D_\mu, D_\nu] = \mathrm{i}F_{\mu\nu}$. The Seeley–DeWitt coefficients now involve traces over both spacetime and internal indices. Writing tr for the internal trace and $\int$ for the spacetime integral:

$\int \mathrm{d}^d x \ \mathrm{tr}\ a_0 = \int \mathrm{d}^d x \ \dim(\text{representation})$. (Volume times the dimension of the field's representation.)

$\int \mathrm{d}^d x \ \mathrm{tr}\ a_1 = \int \mathrm{d}^d x \ \mathrm{tr}\ V$. (As before, with an internal trace.)

$\int \mathrm{d}^d x \ \mathrm{tr}\ a_2 = \int \mathrm{d}^d x \left[\frac{1}{2}\,\mathrm{tr}\, V^2 + \frac{1}{6}\,\mathrm{tr}\,\nabla^2 V + \frac{1}{12}\,\mathrm{tr}(F_{\mu\nu}F^{\mu\nu}) + \text{curvature terms}\right]$.

The crucial term is $\frac{1}{12}\,\mathrm{tr}(F_{\mu\nu}F^{\mu\nu})$: this is the **Yang–Mills action**. Substituting into Schwinger's formula:

$$\frac{1}{2}\,\mathrm{Tr}\log D \supset \frac{1}{(4\pi)^{d/2}} \cdot \frac{\log\Lambda}{24} \int \mathrm{d}^d x \ \mathrm{tr}(F_{\mu\nu}F^{\mu\nu}) + \text{finite}. \tag{12.9}$$

This is the Yang–Mills kinetic term for the gauge field, generated by integrating out charged matter fields. The induced gauge coupling $g_{\mathrm{YM}}^{-2} \propto \log\Lambda$ runs logarithmically with the cutoff, which is the one-loop contribution to **asymptotic freedom** (when the matter contribution is outweighed by the gauge-field self-interaction contribution, the total coupling decreases at short distances).

The parallel to the gravitational case is striking:

	Gravity	**Yang–Mills**
Background field	Metric $g_{\mu\nu}$	Gauge field A_μ^a
Covariant derivative	∇_μ (Christoffel)	$D_\mu = \partial_\mu + \mathrm{i}A_\mu$
Curvature	Riemann $R_{\mu\nu\rho\sigma}$	Field strength $F_{\mu\nu}$
Heat kernel term	$a_1 \supset R$	$a_2 \supset \mathrm{tr}(F^2)$
Generated action	Einstein–Hilbert $\int \sqrt{g}\, R$	Yang–Mills $\int \mathrm{tr}(F^2)$
Divergence	Λ^{d-2} (power)	$\log\Lambda$ (logarithmic)

Both arise from the same mechanism: the Tr log of a matter field on a non-trivial (curved or gauged) background generates an effective action for the background field. The Seeley–DeWitt coefficients determine which geometric or gauge-theoretic terms appear.

Let us derive a_2 in detail. We need the second-order Duhamel expansion. The heat kernel for $D = -\nabla^2 + V$ satisfies

$$e^{-tD} = e^{t\nabla^2} - \int_0^t \mathrm{d}s_1\, e^{(t-s_1)\nabla^2} V\, e^{s_1\nabla^2} + \int_0^t \mathrm{d}s_1 \int_0^{s_1} \mathrm{d}s_2\, e^{(t-s_1)\nabla^2} V\, e^{(s_1-s_2)\nabla^2} V\, e^{s_2\nabla^2} + \cdots$$

The second-order contribution to $K(t,x,x)$ involves two insertions of V at positions y_1 and y_2, connected by free heat kernels:

$$\delta^{(2)}K(t,x,x) = \int_0^t \mathrm{d}s_1 \int_0^{s_1} \mathrm{d}s_2 \int \mathrm{d}^d y_1\ \mathrm{d}^d y_2\ K_0(t-s_1, x, y_1)\, V(y_1)\, K_0(s_1-s_2, y_1, y_2)\, V(y_2)\, K_0(s_2, y_2, x),$$

where $K_0(t,x,y) = (4\pi t)^{-d/2} \exp(-|x-y|^2/(4t))$ is the free heat kernel.

For the coincident-point trace, the leading behavior as $t \to 0$ comes from $y_1 \approx y_2 \approx x$. Expanding $V(y_1)V(y_2) \approx V(x)^2 + O(\nabla V)$ and performing the Gaussian integrals over $y_1 - x$ and $y_2 - x$:

The V^2 term: setting $V(y_1) = V(y_2) = V(x)$ and using the semigroup property $\int K_0(t_1, x, y)K_0(t_2, y, z)\,\mathrm{d}y = K_0(t_1 + t_2, x, z)$, the double integral over s_1, s_2 gives

$$V(x)^2 \int_0^t \mathrm{d}s_1 \int_0^{s_1} \mathrm{d}s_2\; K_0(t, x, x) = \frac{t^2}{2} V(x)^2 \cdot \frac{1}{(4\pi t)^{d/2}},$$

contributing $\frac{1}{2}V(x)^2$ to $a_2(x)$.

The gradient correction: expanding $V(y_1) \approx V(x)+(y_1-x)\cdot\nabla V(x)+\frac{1}{2}(y_1-x)^\top\nabla^2 V(x)(y_1-x)$, the linear term vanishes by symmetry, and the quadratic term produces a correction from the Gaussian integral $\int (y_1 - x)_\mu (y_1 - x)_\nu\, K_0(s, x, y_1)\,\mathrm{d}y_1 = 2s\,\delta_{\mu\nu}$. After careful bookkeeping of the time integrals and the non-commutativity $[\nabla^2, V] \neq 0$ (which generates additional gradient terms when the operators are reordered), the result is $\frac{1}{6}\nabla^2 V(x)$.

Combining:

$$a_2(x) = \frac{1}{2}V(x)^2 + \frac{1}{6}\nabla^2 V(x).$$

The factor $\frac{1}{6}$ from the gradient term is a universal coefficient that appears in the heat kernel expansion on flat space. On a curved Riemannian manifold with Ricci scalar R, the coefficient a_1 acquires an additional term $-\frac{1}{6}R$, and a_2 includes curvature-squared terms (R^2, $R_{\mu\nu}R^{\mu\nu}$, $R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma}$). These are the Seeley–DeWitt coefficients for curved spaces, and they play a central role in quantum gravity and the computation of trace anomalies.

12.6 From Heat Kernel to the Effective Action

Substituting the Seeley–DeWitt expansion into Schwinger's formula, the one-loop effective action $\frac{1}{2}\operatorname{Tr}\log D$ decomposes into local contributions:

$$\frac{1}{2}\operatorname{Tr}\log D = \frac{1}{2(4\pi)^{d/2}}\left[\frac{\Lambda^d}{d/2}\int a_0 + \frac{\Lambda^{d-2}}{(d-2)/2}\int a_1 + \cdots + \log\Lambda \int a_{d/2} + \text{finite}\right], \tag{12.10}$$

where Λ is a UV cutoff. The UV-divergent terms come from the $t \to 0$ region:

$\int a_0\;\mathrm{d}^d x = \int \mathrm{d}^d x$ (volume): coefficient is the **cosmological constant**.

$\int a_1\;\mathrm{d}^d x = \int V(x)\;\mathrm{d}^d x$ (integrated potential): gives **mass renormalization**—the tadpole.

$\int a_2\;\mathrm{d}^d x$ (quadratic in V, gradients): on a curved background, contains $\int \sqrt{g}\,R\;\mathrm{d}^d x$, the **Einstein–Hilbert action**. This is how gravity emerges from the one-loop determinant of matter fields.

This is the precise mechanism by which UV divergences arise: the short-time behavior of the heat kernel generates divergent contributions from low-order Seeley–DeWitt coefficients. Renormalization absorbs these divergences into the coefficients of the corresponding local operators.

12.7 The Massive Scalar in Detail

For $D = -\nabla^2 + m^2$ on flat $\mathbb{R}^d$, the heat kernel at coincident points is

$$K(t, x, x) = \frac{1}{(4\pi t)^{d/2}}\, e^{-m^2 t}.$$

The exponential $e^{-m^2 t}$ comes from $a_n = m^{2n}/n!$ (since $V = m^2$ is constant, the Duhamel expansion exponentiates). The Schwinger integral gives

$$\frac{1}{2}\,\mathrm{Tr}\log(-\nabla^2 + m^2) = \frac{\mathrm{Vol}}{2}\int \frac{d^d k}{(2\pi)^d}\log(k^2+m^2) = \frac{\mathrm{Vol}}{2(4\pi)^{d/2}}\int_0^\infty \frac{dt}{t^{1+d/2}}\,e^{-m^2 t}.$$

The proper-time integral diverges at $t \to 0$ (UV). Introducing a cutoff $t > \epsilon = 1/\Lambda^2$:

$$d = 4: \quad \frac{\mathrm{Vol}}{32\pi^2}\left[\Lambda^4 - m^2\Lambda^2 + \frac{m^4}{2}\left(\log\frac{\Lambda^2}{m^2} - \frac{3}{2}\right) + O(1/\Lambda^2)\right].$$

The three terms correspond to Seeley–DeWitt coefficients: Λ^4 from $a_0 = 1$ (cosmological constant), $\Lambda^2 m^2$ from $a_1 = m^2$ (mass renormalization), $m^4 \log\Lambda$ from $a_2 = m^4/2$ (wave-function/coupling renormalization). These are precisely the divergences that must be absorbed by renormalization in ϕ^4 theory.

Let us walk through the evaluation in $d = 4$ step by step. Starting from

$$\frac{1}{2}\,\mathrm{Tr}\log(-\nabla^2 + m^2) = \frac{\mathrm{Vol}}{2(4\pi)^2}\int_{1/\Lambda^2}^\infty \frac{dt}{t^3}\,e^{-m^2 t},$$

we need to evaluate $I = \int_{1/\Lambda^2}^\infty t^{-3} e^{-m^2 t}\,dt$. Make the substitution $u = m^2 t$:

$$I = m^4 \int_{m^2/\Lambda^2}^\infty u^{-3} e^{-u}\,du = m^4\,\Gamma(-2, m^2/\Lambda^2),$$

where $\Gamma(-2, x)$ is the upper incomplete gamma function. For small $x = m^2/\Lambda^2 \ll 1$ (which is the regime where the cutoff Λ is much larger than the mass m), we use the asymptotic expansion:

$$\Gamma(-2, x) = \frac{1}{2x^2} - \frac{1}{x} + \frac{1}{2}\left(\log\frac{1}{x} + \gamma_E - \frac{3}{2}\right) + O(x),$$

where $\gamma_E \approx 0.5772$ is the Euler–Mascheroni constant. Multiplying by m^4 and substituting $x = m^2/\Lambda^2$:

$$\begin{aligned} I &= m^4\left[\frac{\Lambda^4}{2m^4} - \frac{\Lambda^2}{m^2} + \frac{1}{2}\left(\log\frac{\Lambda^2}{m^2} + \gamma_E - \frac{3}{2}\right) + O\!\left(\frac{m^2}{\Lambda^2}\right)\right] \\ &= \frac{\Lambda^4}{2} - m^2\Lambda^2 + \frac{m^4}{2}\left(\log\frac{\Lambda^2}{m^2} - \frac{3}{2} + \gamma_E\right) + O\!\left(\frac{m^6}{\Lambda^2}\right). \end{aligned}$$

The three divergent terms are precisely the contributions from the Seeley–DeWitt coefficients:

$\Lambda^4/2$: from $a_0 = 1$, this is a quartically divergent constant (the cosmological-constant contribution). It is independent of m and does not affect any observable that is defined as a difference of actions.

$-m^2\Lambda^2$: from $a_1 = m^2$, this is a quadratically divergent mass correction—the tadpole. It shifts the bare mass by an amount that diverges as $\Lambda \to \infty$. Renormalization absorbs this by adjusting $m_0^2(\Lambda)$.

$\frac{m^4}{2}\log(\Lambda^2/m^2)$: from $a_2 = m^4/2$, this is a logarithmically divergent correction to the quartic coupling. It is the one-loop contribution to the running of the coupling constant λ.

Terms of order m^6/Λ^2 and higher are *finite* as $\Lambda \to \infty$: they are the contributions from $a_3, a_4, \ldots$, which produce convergent proper-time integrals. These terms are the predictions of the theory; the divergent terms are absorbed into renormalized parameters.

On a saddle background. For $D = -\nabla^2 + m^2 + \sigma(x)$, the Seeley–DeWitt coefficients become $a_1(x) = m^2 + \sigma(x)$ and $a_2(x) = (m^2 + \sigma)^2/2 + \nabla^2\sigma/6$. Integrating a_2 over space: $\int a_2 \, \mathrm{d}^d x = \int [\frac{1}{2}(m^2 + \sigma)^2 + \frac{1}{6}\nabla^2\sigma] \, \mathrm{d}^d x$. Under periodic boundary conditions, $\int \nabla^2\sigma \, \mathrm{d}^d x = 0$, so the integrated a_2 contributes only $\frac{1}{2}\int (m^2 + \sigma)^2 \, \mathrm{d}^d x$ — a local potential, not a gradient term.

Where, then, does the kinetic energy $(\nabla\sigma)^2$ for the collective field come from? It comes from the *momentum dependence of the polarization bubble*. Recall from Chapter 15 that the collective-field Hessian is $K(q) = D_0^{-1} + \Pi(q)$, and the bubble has a small-q expansion $\Pi(q) = \Pi(0) + Z_\sigma q^2 + O(q^4)$. The coefficient Z_σ is the "wavefunction renormalization" of the collective field—it gives σ a gradient energy $Z_\sigma(\nabla\sigma)^2$ that is generated by the momentum dependence of the microscopic propagator loop, not by the local heat-kernel expansion. The heat kernel captures only the $q = 0$ (local) part of the effective action; the gradient terms require the full momentum structure of the bubble.

12.8 The Lattice Laplacian with Spacing a

On a d-dimensional cubic lattice with spacing a and N^d sites, the lattice Laplacian acts as

$$(-\Delta_{\text{lat}}\phi)(x) = \frac{1}{a^2}\sum_{\mu=1}^{d}\left[2\phi(x) - \phi(x + a\hat{\mu}) - \phi(x - a\hat{\mu})\right].$$

Its Fourier transform is $\tilde{D}(k) = \frac{4}{a^2}\sum_\mu \sin^2(k_\mu a/2) + m^2$, with momenta restricted to the Brillouin zone $|k_\mu| \leq \pi/a$. The lattice provides a natural UV cutoff: $\Lambda = \pi/a$. No modes exist above this momentum. Every integral is finite.

The heat trace on the lattice is an ordinary finite sum:

$$\text{Tr}\, e^{-tD_{\text{lat}}} = \sum_{k\in\text{BZ}} e^{-t[\frac{4}{a^2}\sum_\mu \sin^2(k_\mu a/2) + m^2]}. \tag{12.11}$$

At small t (UV regime), the sum is dominated by high-momentum modes near the Brillouin zone boundary. Converting the sum to an integral (valid for large lattices):

$$\text{Tr}\, e^{-tD_{\text{lat}}} \approx \left(\frac{L}{a}\right)^d \int_{-\pi/a}^{\pi/a} \frac{d^d k}{(2\pi/a)^d}\, e^{-t[\frac{4}{a^2}\sum_\mu \sin^2(k_\mu a/2) + m^2]}.$$

At small t, the dominant contribution comes from modes near $k = 0$ (where $\sin^2(ka/2) \approx k^2a^2/4$) and from modes near the Brillouin zone boundary $k = \pi/a$. (Note: unlike the fermionic lattice Dirac operator, the scalar lattice Laplacian does not suffer from a doubling problem—the modes near $k = \pi/a$ have energies of order $1/a^2$ and decouple in the continuum limit.) The expansion reproduces the continuum result with corrections:

$$\text{Tr}\, e^{-tD_{\text{lat}}} = \frac{\text{Vol}}{(4\pi t)^{d/2}}\, e^{-m^2 t}\left[1 + c_1\frac{t}{a^2} + c_2\frac{t^2}{a^4} + \cdots\right], \tag{12.12}$$

where the corrections $c_n(t/a^2)^n$ arise from the difference between $\sin^2(ka/2)$ and $(ka/2)^2$. These are *lattice artifacts*: they vanish in the continuum limit $a \to 0$ at fixed t.

The Tr log on the lattice is therefore

$$\frac{1}{2}\,\text{Tr}\log D_{\text{lat}} = \frac{1}{2}\sum_{k\in\text{BZ}} \log\left[\frac{4}{a^2}\sum_\mu \sin^2\frac{k_\mu a}{2} + m^2\right].$$

This is a finite sum of logarithms—no divergences, no regularization needed. The lattice spacing a simultaneously provides the UV cutoff ($\Lambda = \pi/a$) and discretizes the $\operatorname{Tr}\log$ into a computable sum. When we take $a \to 0$ with the bare parameters $\theta(a)$ tuned according to the renormalization prescription of Chapter 37, the lattice artifacts $O(a^2)$ vanish and the physical predictions converge to the continuum limit.

12.9 Reliability of the Heat-Kernel Coefficients: What Can Be Trusted

The heat-kernel expansion generates the *correct operator structure* of the effective action: every local operator allowed by the emergent symmetries appears. But the *numerical coefficients* of these operators have very different levels of reliability. Understanding this hierarchy is essential for anyone using the heat kernel in practice, and it illuminates one of the deepest puzzles in theoretical physics: the cosmological constant problem.

The microscopic lattice versus the continuum approximation

Recall that every continuum functional integral in this book is a shorthand for a lattice integral (Chapter 1, Part XI). The continuum Laplace-type operator $D = -\nabla^2 + V$ approximates the lattice Laplacian L_a only for modes with wavelength much longer than the lattice spacing: $ka \ll 1$. Near the Brillouin zone boundary ($ka \sim 1$), the continuum description fails.

This distinction is directly reflected in the heat kernel. The *lattice* heat trace has the strict short-time expansion

$$\operatorname{Tr} e^{-sL_a} = N - s \operatorname{Tr} L_a + \frac{s^2}{2} \operatorname{Tr} L_a^2 + \cdots, \qquad s \to 0^+, \tag{12.13}$$

which is analytic in s, beginning with the number of sites N. The *continuum* heat kernel has the asymptotic behavior $(4\pi s)^{-d/2} \sum a_n s^n$, which is singular at $s = 0$. The continuum form cannot describe the strict $s \to 0$ limit of the lattice; it is an **intermediate asymptotic**, valid in the window

$$\boxed{a^2 \ll s \ll L_{\text{curv}}^2,} \tag{12.14}$$

where a is the lattice spacing and L_{curv} is the scale on which the geometry varies. The diffusion length $\sqrt{s}$ must be large compared with a (so the continuum approximation is valid) but small compared with the curvature scale (so the local expansion is valid).

UV weighting and the reliability hierarchy

The proper-time representation of the one-loop effective action is

$$\Gamma_{\text{1-loop}} = \frac{1}{2} \int_0^\infty \frac{ds}{s} \operatorname{Tr} e^{-sD},$$

with a regulator understood. Substituting the Seeley–DeWitt expansion and integrating term by term:

$$\Gamma_{\text{local}} \sim \int \mathrm{d}^d x \sqrt{g} \left[c_0 \Lambda^d A_0 + c_1 \Lambda^{d-2} A_1 + c_2 \log(\Lambda/\mu) A_2 + \cdots \right].$$

The key observation is that higher powers of Λ weight the integral more heavily toward small s (large momenta), which is precisely the regime where the continuum approximation to the lattice is *least accurate*. This creates a hierarchy of reliability:

The A_0 term (cosmological constant): proportional to Λ^d. This receives its dominant contribution from the smallest proper times ($s \sim 1/\Lambda^2$), corresponding to the highest-momentum modes. But these are exactly the modes for which the curved continuum operator D is least faithful to the microscopic lattice. The coefficient depends on the exact lattice dispersion relation, the density of microscopic degrees of freedom, the Brillouin-zone measure, and the normalization of the emergent metric—none of which are determined by the long-wavelength Laplacian alone.

The A_1 term (Einstein–Hilbert): proportional to Λ^{d-2}. Still UV-sensitive, but less so than A_0. More importantly, the Ricci scalar R measures the *response* of the spectrum to a slowly varying deformation of the geometry—a response that can be probed within the regime where the continuum description is valid. The coefficient can be determined by a matching calculation: compare the lattice spectral response on slowly varying backgrounds with the continuum prediction.

The A_2 term (higher-derivative gravity / Yang–Mills): proportional to $\log \Lambda$ (in $d = 4$). Only logarithmically sensitive to the UV, making its coefficient much more stable under changes of the microscopic theory.

Finite (non-divergent) terms: independent of Λ. These are the most universal predictions of the continuum theory, determined entirely by the infrared physics.

The cosmological constant problem

This hierarchy illuminates why the cosmological constant is so problematic. In $d = 4$, a naive estimate gives $\rho_\Lambda \sim \Lambda^4/(16\pi^2)$. If Λ is the Planck scale ($\sim 10^{19}$ GeV), this gives $\rho_\Lambda \sim 10^{76}$ GeV^4, while the observed cosmological constant is $\rho_\Lambda^{\text{obs}} \sim 10^{-47}$ GeV^4—a discrepancy of 123 orders of magnitude.

From the perspective of this book, the resolution is clear in principle: the naive estimate is the *least trustworthy* output of the continuum approximation, because it is dominated by the UV regime where the approximation fails. The actual cosmological constant is a *renormalized parameter*:

$$\rho_\Lambda = \rho_\Lambda^{\text{micro}} + \rho_\Lambda^{\text{IR}} + \rho_\Lambda^{\text{ct}}, \tag{12.15}$$

where $\rho_\Lambda^{\text{micro}}$ is the microscopic (lattice-scale) contribution, ρ_Λ^{IR} is the contribution from the long-wavelength continuum theory, and ρ_Λ^{ct} is a counterterm (or, equivalently, the bare cosmological constant). The heat-kernel calculation gives ρ_Λ^{IR}, but the physical value is the sum of all three terms, and there is no reason for the sum to be close to any individual term.

The curvature coefficient (Newton's constant) faces a similar problem but is somewhat more tractable, because it can be determined by matching the spectral response of the lattice to a slowly varying background metric. The matching procedure is well-defined within the continuum regime $ka \ll 1$ and does not require extrapolation to the UV.

The sharp summary

The UV–IR inversion: a structural analysis

The hierarchy of UV sensitivity established above has a remarkable property: it is *exactly inverted* relative to the hierarchy of IR relevance.

Consider an operator $\mathcal{O}_n$ with $2n$ derivatives acting on the metric (or gauge field). In d spacetime dimensions:

Operator	Derivatives	UV sensitivity	IR scaling	RG relevance
$\int \sqrt{g}$ (cosmological)	0	Λ^d	k^0	most relevant
$\int \sqrt{g}\, R$ (Einstein–Hilbert)	2	Λ^{d-2}	k^2	relevant ($d > 2$)
$\int \sqrt{g}\, R^2$ (higher-deriv.)	4	Λ^{d-4}	k^4	marginal ($d = 4$)
$\int \sqrt{g}\, R^3$	6	Λ^{d-6}	k^6	irrelevant ($d < 6$)

The pattern is exact and follows from dimensional analysis alone. An operator with $2n$ derivatives has engineering dimension $d - 2n$ (in mass units). To appear in the action (which is dimensionless), its coefficient must have dimension $2n - d$, which is compensated by Λ^{d-2n} from the UV integral. Meanwhile, at long wavelengths ($k \to 0$), the contribution of this operator to any amplitude scales as k^{2n} relative to the leading (zero-derivative) term.

The UV sensitivity Λ^{d-2n} *decreases* with n, while the IR suppression k^{2n} *increases* with n. The crossover is exact: **the operator most contaminated by UV physics is the one that dominates the IR.**

This is not a coincidence; it is a direct consequence of the structure of the derivative expansion. The operators with the fewest derivatives are the most relevant in the IR precisely because they have the highest engineering dimension, and operators with the highest engineering dimension have the most UV-divergent coefficients precisely because more powers of Λ are needed to compensate their dimension. The UV–IR inversion is built into the dimensional analysis of effective field theory.

Why this inversion matters: the cosmological constant problem

The cosmological constant problem is the sharpest manifestation of the UV–IR inversion. In $d = 4$:

The *operator* $\int \sqrt{g}\, \mathrm{d}^4x$ is the most relevant term in the gravitational effective action. At distances much larger than $1/\sqrt{\rho_\Lambda}$, it dominates over the Einstein–Hilbert and higher-derivative terms. Observationally, it determines the late-time expansion of the universe.

The *coefficient* ρ_Λ receives contributions $\sim \Lambda^4$ from every field in the theory. If $\Lambda \sim M_{\text{Planck}} \sim 10^{19}$ GeV, the one-loop estimate gives $\rho_\Lambda \sim 10^{76}$ GeV^4, while the observed value is $\rho_\Lambda^{\text{obs}} \sim 10^{-47}$ GeV^4—a discrepancy of 123 orders of magnitude.

But from the lattice perspective developed in this book, the discrepancy is a *failure of predictivity*, not of physics. The Λ^4 estimate extrapolates the continuum heat kernel into the regime $s \lesssim a^2$ where it is not valid. The actual coefficient is a renormalized parameter that includes microscopic (lattice-scale) contributions not captured by the continuum approximation:

$$\rho_\Lambda = \rho_\Lambda^{\text{micro}} + \rho_\Lambda^{\text{continuum}} + \rho_\Lambda^{\text{ct}}. \tag{12.16}$$

The heat-kernel calculation gives only $\rho_\Lambda^{\text{continuum}}$, which is large and UV-sensitive. But the physical ρ_Λ is the sum of all three terms, and there is no reason for the sum to be close to any individual term. The smallness of the observed ρ_Λ relative to M_{Planck}^4 may reflect a near-cancellation between microscopic and continuum contributions—or it may reflect physics entirely outside the framework of perturbative quantum field theory.

The condensed-matter perspective

In condensed matter, the UV–IR inversion is present but harmless, for an illuminating reason.

The "cosmological constant" of a condensed-matter system is the ground-state energy density E_0/V. This quantity is indeed dominated by UV modes: every filled state below the Fermi energy contributes, and most of the energy comes from the deepest, most UV states. The quantity is of order $n\varepsilon_F \sim \Lambda^d$ (where n is the particle density and ε_F the Fermi energy), and its precise value depends sensitively on the lattice structure, the atomic binding energies, and other microscopic details.

But E_0/V is *not a measurable quantity* in condensed matter. All physical observables depend on *energy differences*: excitation energies, free-energy differences between phases, response functions. These correspond to the higher Seeley–DeWitt coefficients (A_1, A_2, ...), which are less UV-sensitive and more reliably computed from the continuum theory.

The reason the cosmological constant problem is a problem in gravity, but not in condensed matter, is that **gravity couples to absolute energy density**. The Einstein equations $G_{\mu\nu} = 8\pi G\, T_{\mu\nu}$ contain $T_{00} = \rho$, the total energy density, not a relative energy. This makes the most UV-sensitive quantity (the absolute vacuum energy) also the most physically consequential—a situation that does not arise in any other branch of physics.

Does every mode contribute equally to the UV-sensitive coefficient?

A more quantitative version of the question: in the Wilsonian RG, each momentum shell $[k, k + dk]$ contributes to the running of the cosmological constant. The contribution from a single shell is

$$\delta\rho_\Lambda = \frac{S_d\, k^{d-1}\, \mathrm{d}k}{2(2\pi)^d} \sqrt{k^2 + m^2} \quad \text{(bosonic, zero-point energy per mode)},$$

where S_d is the area of the unit sphere. Summing over all shells up to Λ:

$$\rho_\Lambda = \int_0^\Lambda \frac{S_d\, k^{d-1}}{2(2\pi)^d} \sqrt{k^2 + m^2}\, \mathrm{d}k \approx \frac{S_d}{2(2\pi)^d} \cdot \frac{\Lambda^{d+1}}{d+1} \quad (\Lambda \gg m).$$

The integrand $k^{d-1}\sqrt{k^2 + m^2} \approx k^d$ grows with k: **each UV shell contributes more than the previous one**. The total is dominated by the last shell (near Λ), which is exactly the shell where the continuum approximation to the lattice is least accurate.

For the Einstein–Hilbert coefficient ($A_1 \supset R$), the analogous shell contribution is $\delta M_{\text{eff}}^{d-2} \sim k^{d-3}\,\mathrm{d}k$. In $d = 4$, this gives $\delta M_{\text{eff}}^2 \sim k\, dk$, still UV-dominated but less severely.

For the A_2 coefficient ($\sim \log \Lambda$), the shell contribution is $\delta c_{R^2} \sim k^{d-5}\,\mathrm{d}k$. In $d = 4$, this gives $\delta c_{R^2} \sim dk/k$, a *logarithmic* accumulation—each shell contributes equally. This is why $\log \Lambda$ divergences are "universal": they receive equal contributions from all scales, not just the UV.

For finite terms, the integrand decays as k^{d-2n-1} with $2n > d$, and the integral converges: the UV shells contribute less than the IR shells. These terms are genuinely determined by the infrared.

The hierarchy is therefore:

Coefficient	**Shell weight**	**Dominated by**	**Predictivity**
ρ_Λ (A_0)	k^d (growing)	UV modes	Least predictable
M_{eff}^{d-2} (A_1)	k^{d-2} (growing)	UV modes (less so)	Low
c_{R^2} (A_2, $d = 4$)	k^{-1} (flat)	All scales equally	Moderate (universal log)
Finite terms	$k^{<0}$ (falling)	IR modes	Most predictable

The UV–IR inversion is exact: the operator that dominates the IR (k^0, most relevant) is the one whose coefficient is most dominated by the UV (k^d, most divergent). This is a structural feature of effective field theory, not an accident.

Exceptions and qualifications

The UV–IR inversion, while structurally robust, has important exceptions that clarify its scope:

Symmetry protection. If a symmetry relates the contributions of different fields, the UV sensitivity can be reduced. Supersymmetry pairs every boson with a fermion of equal mass, and since bosonic and fermionic zero-point energies have opposite sign, the leading Λ^d contribution to ρ_Λ cancels. With exact supersymmetry, $\rho_\Lambda = 0$ perturbatively. With broken supersymmetry at scale M_{SUSY}, the cosmological constant is $\rho_\Lambda \sim M_{\text{SUSY}}^4$ rather than Λ^4—still large, but parametrically smaller. This illustrates that the UV–IR inversion can be tamed by symmetry, even if it cannot be eliminated.

Topological terms. The theta term in QCD, $\theta \int \text{tr}(F\tilde{F})$, has zero derivatives of the metric (it involves only the gauge field) and dimension 4, yet it receives *no* perturbative UV divergence because $F\tilde{F} = \partial_\mu K^\mu$ is a total derivative. Its coefficient is a free parameter of the theory but is not UV-sensitive. Topological terms evade the UV–IR inversion because they do not appear in the local Seeley–DeWitt expansion.

Anomalies. The trace anomaly generates a contribution to the effective action of the form $c \int \sqrt{g}\,(R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} - \cdots)$, where c is a dimension-dependent coefficient computable from the field content. This coefficient is *exact* (receives no corrections from any scale) and UV-insensitive, despite being associated with a dimension-4 operator. Anomalous contributions are protected by topological constraints and are exceptions to the general hierarchy.

Nonlocal IR terms. Beyond the local Seeley–DeWitt expansion, the Tr log produces nonlocal terms that cannot be written as integrals of local densities. For example, integrating out a massless scalar on a curved background generates $R\,\Box^{-1}R$, where $\Box^{-1}$ is the inverse d'Alembertian—a manifestly long-range, IR-dominated contribution. Such terms are determined by the mass spectrum and couplings of light fields, not by UV details, and they are *more universal* than the local power-divergent terms.

These nonlocal terms are precisely the terms that carry physical information about collective-mode propagation. In the language of this book: the local part of S_{eff} (the saddle, the gap, the Seeley–DeWitt coefficients) is UV-sensitive and requires renormalization. The nonlocal part (the propagator $D = K^{-1}$, the dispersion relations, the spectral functions) is determined by the momentum-dependent polarization $\Pi(q)$ and is the genuinely predictive output of the theory.

Massive fields. If all fields have mass m, the UV contributions to ρ_Λ are cut off at $k \sim m$ rather than $k \sim \Lambda$, and the coefficient becomes $\rho_\Lambda \sim m^d$ (the Euler–Heisenberg result). The UV–IR inversion is then between m and the IR scale, not between Λ and the IR scale. This is the "decoupling" of heavy fields: their contribution to the cosmological constant is set by their mass, not by the cutoff. The cosmological constant problem then becomes: why don't the known massive particles (with masses up to $m_t \sim 170$ GeV) generate a cosmological constant of order $m_t^4 \sim 10^9$ GeV^4, still far larger than observed?

Subtraction of a flat reference

A practical strategy for reducing the UV sensitivity is to subtract a flat-background reference:

$$\Delta\Gamma[g] = \Gamma[g] - \Gamma[g_{\text{flat}}].$$

If the comparison is made at fixed physical volume with an identical number of microscopic degrees of freedom, the leading A_0 contribution cancels (since $\int \sqrt{g}\;\mathrm{d}^dx = \int \mathrm{d}^dx = \text{Vol}$ for both), making the curvature-dependent terms (A_1, A_2, ...) more reliable. This is the gravitational analogue of measuring energy differences rather than absolute energies—the condensed-matter strategy that eliminates the cosmological constant problem.

However, the cancellation is not automatic. The emergent volume element may itself depend on the microscopic deformation (if the lattice is distorted, the "physical volume" as seen by the continuum theory may differ from the "coordinate volume" as counted by lattice sites). Fixing coordinate volume, physical volume, or lattice-site number are different prescriptions, and a residual cosmological term can survive. The subtraction procedure must be defined at the microscopic level before one can claim that the leading UV contribution has been removed.

Summary: what can and cannot be trusted

The heat-kernel expansion identifies the correct operator structure of the effective action. The numerical coefficients have a hierarchy of reliability that is exactly inverted relative to their IR importance:

Most robust (structure): every local operator allowed by the emergent symmetries appears. The *form* of the effective action—cosmological term, Einstein–Hilbert, Yang–Mills, higher derivatives—is dictated by symmetry alone.

Most robust (coefficients): finite, nonlocal terms determined by light-field propagation (e.g., $R\Box^{-1}R$). These are the analogues of the collective-mode propagator $D(q)$ in the condensed-matter setting.

Conditionally robust: logarithmically divergent coefficients (A_2 in $d = 4$). These receive equal contributions from all scales and are "universal" in the sense that they depend on the field content but not on the details of the UV completion.

Matchable but UV-sensitive: power-divergent coefficients with Λ^{d-2n} for $n \geq 1$ (A_1: Newton's constant). These can be determined by spectral matching to the lattice within the continuum regime.

Least trustworthy: the Λ^d coefficient (A_0: cosmological constant). Dominated by the UV regime where the continuum approximation fails. Its value is a renormalized parameter that must be determined by measurement.

The UV–IR inversion—that the most IR-relevant operator has the least UV-predictable coefficient—is a structural feature of dimensional analysis in effective field theory. It is not a bug to be fixed but a fact to be respected.

12.10 Entropy, Conditional Entropy, and Mutual Information

The Tr log that appears throughout this book has a direct information-theoretic interpretation that connects to the entropy of Gaussian fields and to variational inference in machine learning.

Gaussian entropy

For a Gaussian random field ϕ with covariance C on a lattice of V sites, the differential entropy is

$$H(\phi) = \frac{V}{2}\log(2\pi e) + \frac{1}{2}\log\det C = \frac{V}{2}\log(2\pi e) + \frac{1}{2}\operatorname{Tr}\log C. \tag{12.17}$$

The $\frac{1}{2}\operatorname{Tr}\log C$ term is the *same* Tr log that we encounter throughout: writing $C = D^{-1}$ (precision = inverse covariance), $\frac{1}{2}\operatorname{Tr}\log C = -\frac{1}{2}\operatorname{Tr}\log D$. The entropy of a Gaussian field is (minus) the Tr log of its precision.

Conditional entropy as partial access

In the HST framework, we have a joint distribution $p(\phi, \sigma)$. The **conditional entropy** $H(\phi|\sigma)$ measures the remaining uncertainty about ϕ when σ is observed:

$$H(\phi|\sigma) = -\int p(\phi, \sigma) \log p(\phi|\sigma)\ \mathrm{d}\phi\ \mathrm{d}\sigma.$$

For the Gaussian–Gaussian model where $p(\phi|\sigma)$ is Gaussian with precision $A = A(\sigma)$, the conditional entropy is $H(\phi|\sigma) = \frac{V}{2}\log(2\pi e) - \frac{1}{2}\mathbb{E}_\sigma[\operatorname{Tr}\log A(\sigma)]$. At the saddle $\bar{\sigma}$:

$$H(\phi|\bar{\sigma}) = \frac{V}{2}\log(2\pi e) - \frac{1}{2}\operatorname{Tr}\log M(\bar{\sigma}). \tag{12.18}$$

The $\operatorname{Tr}\log M(\bar{\sigma})$ in the effective action is (up to constants) the *negative conditional entropy of the microscopic field given the collective field at its saddle value.* Integrating out ϕ reduces the entropy from $H(\phi)$ to $H(\phi|\bar{\sigma})$: the collective field σ captures part of the information about ϕ, and the remaining uncertainty is encoded in the fluctuation determinant.

Mutual information

The **mutual information** between ϕ and σ is

$$I(\phi;\sigma) = H(\phi) - H(\phi|\sigma) = \frac{1}{2}\operatorname{Tr}\log C_\phi - \frac{1}{2}\mathbb{E}_\sigma[\operatorname{Tr}\log C_{\phi|\sigma}] = \frac{1}{2}\operatorname{Tr}\log\frac{C_\phi}{C_{\phi|\sigma}}.$$

For the Gaussian case, $C_\phi = (A - WB^{-1}W^\top)^{-1}$ (marginal covariance, from the Schur complement) and $C_{\phi|\sigma} = A^{-1}$ (conditional covariance). The mutual information measures how much the marginalisation changes the effective covariance—it is a scalar summary of the Schur complement correction.

In the variational inference language: the ELBO (evidence lower bound) used in variational autoencoders is $\mathrm{ELBO} = \mathbb{E}_q[\log p(\phi|\sigma)] - \mathrm{KL}(q(\sigma)\|p(\sigma))$, which decomposes the $\log p(\phi)$ (our negative effective action $-S_{\text{eff}}$) into a reconstruction term and a complexity penalty. The Tr log appears in both terms.

The ELBO and the one-loop effective action

The connection between Tr log and the ELBO can be made fully explicit. In variational inference, we approximate the posterior $p(\sigma|\text{data})$ by a variational distribution $q(\sigma)$, and the evidence satisfies

$$\log p(\text{data}) = \underbrace{\mathbb{E}_q[\log p(\text{data}, \sigma)] - \mathbb{E}_q[\log q(\sigma)]}_{\text{ELBO}} + D_{\text{KL}}(q\|p(\sigma|\text{data})).$$

Choosing $q(\sigma) = \mathcal{N}(\bar{\sigma}, K^{-1})$ (a Gaussian centered at $\bar{\sigma}$ with precision K), the ELBO becomes

$$\begin{aligned} \mathrm{ELBO} &= \mathbb{E}_q[\log p(\text{data}, \sigma)] + \frac{1}{2}\log\det K^{-1} + \frac{d}{2}(1 + \log 2\pi) \\ &= \mathbb{E}_q[-S_{\text{eff}}(\sigma)] - \frac{1}{2}\operatorname{Tr}\log K + \text{const.} \end{aligned}$$

Expanding $S_{\text{eff}}(\sigma)$ to second order around $\bar{\sigma}$ (and noting that the linear term vanishes if $\bar{\sigma}$ is the saddle, and the second-order term gives $\frac{1}{2}\operatorname{Tr}(K_{\text{true}}K^{-1})$ where K_{true} is the Hessian of S_{eff}), the ELBO is maximized when $K = K_{\text{true}}$ (the Hessian at the saddle) and $\bar{\sigma}$ is the MAP estimate.

At this optimal point, the ELBO equals

$$\mathrm{ELBO}_{\mathrm{max}} = -S_{\mathrm{eff}}(\bar{\sigma}) - \frac{1}{2}\operatorname{Tr}\log K_{\mathrm{true}} + \frac{d}{2}(1+\log 2\pi),$$

which is exactly the Laplace approximation to $\log Z_\sigma = -S_{\mathrm{eff}}(\bar{\sigma}) - \frac{1}{2}\operatorname{Tr}\log K + \mathrm{const}$ from Chapter 8.

The upshot: the RPA (Gaussian fluctuation theory for the collective field) is equivalent to Laplace inference applied to the posterior of the collective field. The saddle $\bar{\sigma}$ is the MAP estimate. The Hessian K is the Laplace precision (which coincides with the optimal variational precision in the large-N limit). The $\operatorname{Tr}\log K$ correction is the Occam factor (model complexity penalty) in the Bayesian language, and the one-loop determinant in the physics language. These are the same object.

As noted in Chapter 19, the Laplace approximation is not identical to the optimal Gaussian variational approximation at finite N. But the correspondence is exact in the leading-order large-N (or large-volume) limit, which is the same regime where the saddle-point approximation is justified. For the target audience of this book, the essential message is: RPA and Laplace inference are the same mathematical operation, applied to the collective-field posterior.

12.11 The Replica Method

The problem: computing $\langle \log Z \rangle$

In disordered systems (spin glasses, random matrices, neural networks with random weights), the physically relevant quantity is not the partition function Z itself but its *logarithm* $\log Z$—the free energy—averaged over disorder:

$$[\log Z]_{\mathrm{dis}} = \int P(J) \log Z(J)\, \mathrm{d}J,$$

where J denotes the random couplings and $P(J)$ their distribution. The brackets $[\cdot]_{\mathrm{dis}}$ denote the disorder average.

The difficulty is that $\log Z$ is a nonlinear function of Z, and averaging a nonlinear function is much harder than averaging a linear one. The "annealed" average $[\log Z]_{\mathrm{dis}} \le \log[Z]_{\mathrm{dis}}$ (by Jensen's inequality) is easy to compute but overestimates the free energy.

The replica identity

The **replica method** uses the identity

$$\log Z = \lim_{n\to 0} \frac{Z^n - 1}{n}, \tag{12.19}$$

which follows from $Z^n = e^{n\log Z} \approx 1 + n\log Z$ for small n. The key insight: for *integer* n, Z^n is the partition function of n *independent copies* (replicas) of the system:

$$Z^n = \int \prod_{a=1}^{n} \mathrm{d}\phi^{(a)}\; e^{-\sum_{a=1}^{n} S[\phi^{(a)};J]}.$$

The disorder average $[Z^n]_{\mathrm{dis}}$ can be performed *before* the field integral (since the replicas share the same J), typically producing a coupling between replicas:

$$[Z^n]_{\mathrm{dis}} = \int \prod_a \mathrm{d}\phi^{(a)}\; e^{-S_{\mathrm{eff}}[\phi^{(1)},\ldots,\phi^{(n)}]},$$

where S_{eff} now couples different replicas through the disorder average. The quenched free energy is then $[\log Z]_{\text{dis}} = \lim_{n\to 0}([Z^n] - 1)/n$.

Connection to the book's framework

The replica method connects to our framework in several ways. The n replicas are analogous to the N copies in the large-N expansion (Chapter 29)—both introduce multiple copies to simplify the problem. The disorder-averaged S_{eff} generates effective interactions between replicas, just as the HST generates effective interactions between visible variables (Chapter 5). The saddle-point analysis of the replicated system produces order parameters (the overlap matrix $Q_{ab} = \langle \phi^{(a)}\phi^{(b)} \rangle$) that play the role of the collective field σ.

The analytic continuation $n \to 0$ is the mathematically uncontrolled step—it assumes that the replica partition function, computed for integer n, can be analytically continued to $n = 0$. This assumption can fail (and does, in spin glasses with "replica symmetry breaking," where the saddle-point structure for integer n does not have a unique analytic continuation). Nevertheless, the replica method remains one of the most powerful tools for disordered systems, and recent work has placed it on firmer mathematical footing through connections to random matrix theory and the cavity method.

In the machine learning context, the replica method is used to analyze the average-case performance of learning algorithms: the "training loss" is a quenched quantity (averaged over random training data), and the replica method computes its typical value. The order parameters Q_{ab} describe the overlap between different solutions (different trained networks), and replica symmetry breaking corresponds to a rugged loss landscape with many qualitatively different local minima.

12.12 Connection to the Laplace Approximation and Bayesian Model Selection

In Bayesian inference, the Laplace approximation gives

$$\log p(\text{data}) \approx \log p(\text{data}|\hat{\theta}) - \frac{1}{2}\log\det H + \frac{d}{2}\log(2\pi),$$

where H is the Hessian of the negative log-posterior at the MAP estimate $\hat{\theta}$. The term $\frac{1}{2}\log\det H = \frac{1}{2}\operatorname{Tr}\log H$ is the **Occam factor**—the penalty for model complexity.

The heat kernel expansion of $\operatorname{Tr}\log H$ decomposes this Occam factor into contributions from different curvature scales of the log-posterior landscape. In Watanabe's **singular learning theory**, the leading behavior of the Bayesian evidence near singular models is controlled by the real log-canonical threshold λ, which generalizes the Seeley–DeWitt coefficients to the case where the posterior has degenerate (non-isolated) critical points. This is directly relevant to understanding the generalization properties of overparameterized neural networks, where the loss landscape has continuous families of minima.

Part IV

Saddles, Mean Fields, Gaps, and Gap Equations

The effective action $S_{\text{eff}}(\sigma)$, derived in Part III by integrating out Gaussian microscopic variables, is in general a complicated nonlinear functional of the collective field σ. Evaluating the remaining integral $\int \mathcal{D}\sigma\, e^{-S_{\text{eff}}(\sigma)}$ exactly is typically as difficult as the original problem. The simplest approximation is to replace the integral by the contribution from the most probable configuration—the mode or saddle point.

This part develops the saddle-point approximation for the collective field, derives the resulting self-consistency equations (gap equations), and carefully explains what the word "gap" means.

Chapter 13

The Saddle Point as a Mode of a Marginal Distribution

13.1 Finding the Mode

Given the effective action $S_{\text{eff}}(\sigma)$ from equation (8.2) or (9.2), we seek the configuration $\bar{\sigma}$ that minimizes S_{eff}—equivalently, the configuration that maximizes the marginal density $e^{-S_{\text{eff}}(\sigma)}$. This is the mode of the marginal distribution of σ, and it satisfies the stationarity condition

$$\boxed{\left.\frac{\delta S_{\text{eff}}}{\delta \sigma(x)}\right|_{\bar{\sigma}} = 0.} \tag{13.1}$$

In finite dimensions, this is simply $\nabla_\sigma S_{\text{eff}}(\bar{\sigma}) = 0$, the condition for a critical point. If S_{eff} is convex, the critical point is a global minimum and is unique. If S_{eff} has multiple minima, the dominant saddle is (in most contexts) the global minimum, though transitions between minima play an important role in the physics of phase transitions.

13.2 Mode, Mean, and MAP

The saddle $\bar{\sigma}$ is the mode of the marginal distribution $p(\sigma) \propto e^{-S_{\text{eff}}(\sigma)}$. In Bayesian statistics, this is the maximum a posteriori (MAP) estimator. It is important to recognize that the mode is *not* the same as the mean:

$$\bar{\sigma} \neq \mathbb{E}[\sigma]$$

in general. The mode and the mean coincide for symmetric, unimodal distributions (such as a Gaussian), but they differ whenever the distribution is skewed or multimodal, or when the curvature is not uniform. The two are close when the distribution is sharply peaked—that is, when fluctuations around the saddle are small—but the distinction matters in principle and sometimes in practice.

Physicists use several terms for $\bar{\sigma}$ and the approximation built upon it. **Saddle point** refers to the critical point of the action (the term originates from the method of steepest descent in complex analysis, where the critical point of the exponent is a saddle in the complex plane, even when the real-variable critical point is a minimum). **Mean field** is used because in many applications $\bar{\sigma}$ can be interpreted as a self-consistently determined average field seen by each microscopic variable. **Classical background** or **classical solution** is used because in the path-integral formulation of quantum mechanics, the saddle-point approximation corresponds to the classical limit $\hbar \to 0$. **Order parameter** is used when $\bar{\sigma}$ distinguishes between different phases.

All of these refer to the same mathematical object: the mode of a marginal density. The different names reflect different physical contexts, not different mathematical operations.

13.3 When Is the Saddle Approximation Good?

The saddle-point approximation replaces the full integral $\int \mathcal{D}\sigma\, e^{-S_{\text{eff}}(\sigma)}$ by the integrand evaluated at its maximum: $e^{-S_{\text{eff}}(\bar{\sigma})}$. This is a good approximation when the distribution of σ is sharply concentrated around $\bar{\sigma}$, so that contributions from configurations far from the saddle are negligible.

Sharp concentration occurs when S_{eff} has a large overall prefactor. The most common mechanism is a large number of microscopic degrees of freedom. If there are N microscopic variables and S_{eff} scales as N while σ is an $O(1)$ collective variable, then the width of the σ distribution scales as $1/\sqrt{N}$, and the saddle-point approximation becomes exact as $N \to \infty$.

This is precisely the setting of the large-N limit, which we will encounter in Chapter 27 with the $O(N)$ model. More generally, the saddle approximation is the leading term in a systematic expansion. The next correction is the Gaussian fluctuation determinant (the $\frac{1}{2}\operatorname{Tr}\log$ of the Hessian), and further corrections come from higher-order terms in the Taylor expansion around the saddle.

Quantitative example: accuracy of the saddle approximation

To develop quantitative intuition, consider the one-dimensional integral

$$Z(N) = \int_{-\infty}^{\infty} \exp\left(-N\left[\frac{x^2}{2} + \frac{x^4}{4}\right]\right) \mathrm{d}x.$$

The action $f(x) = x^2/2 + x^4/4$ has a unique minimum at $x_* = 0$ with $f(0) = 0$ and $f''(0) = 1$. The Laplace approximation gives $Z_{\text{Laplace}}(N) = \sqrt{2\pi/N}$.

The exact integral can be computed in terms of a parabolic cylinder function, but let us instead compare numerically. The ratio $Z_{\text{exact}}/Z_{\text{Laplace}}$ measures the accuracy of the saddle:

N	Z_{exact}	Z_{Laplace}	Ratio
1	1.812	2.507	0.723
5	0.939	1.121	0.838
20	0.524	0.560	0.935
100	0.245	0.251	0.977

At $N = 1$, the saddle approximation is off by 28%. At $N = 20$, it is accurate to 7%. At $N = 100$, to 2%. The corrections are of order $1/N$, coming from the quartic term $\frac{1}{4!}f^{(4)}(0)\eta^4 = \frac{3}{N}\eta^4$ that the Laplace approximation neglects.

In a field theory on a lattice with V sites, the role of N is played by the number of sites V (or, for the $O(N)$ model, by N). The saddle approximation becomes exact as $V \to \infty$ (or $N \to \infty$), with the first correction being the Gaussian fluctuation determinant $\frac{1}{2}\operatorname{Tr}\log K$, which is itself of order $\log V$ (or $\log N$). For practical purposes: the saddle is a reasonable starting point whenever the effective number of degrees of freedom contributing to the collective-field integral is larger than ~ 10, and it is quantitatively reliable for $N \gtrsim 50$.

The next correction beyond the saddle (the one-loop determinant) improves the accuracy substantially. Including $\frac{1}{2}\log\det(f''(0)) = 0$ (in this case), the one-loop result is the same as the saddle. But if we include the $1/N$ correction from the quartic perturbation of the Gaussian, we get $Z_{\text{1-loop}} = Z_{\text{Laplace}} \cdot (1 - 3/(4N) + \cdots)$, which matches the exact result to $O(1/N^2)$ and brings the $N = 5$ error from 16% to under 3%.

13.4 Two Saddles, Two Classical Limits

There is a profound distinction between the saddle of the Schwinger effective action $\Gamma[\phi_{\text{cl}}]$ and the saddle of the HST effective action $S_{\text{eff}}[\sigma]$. Both saddle points describe macroscopic, classical behavior—but they are the classical behavior of *different things*.

The saddle of $\Gamma[\phi_{\text{cl}}]$ is the physical mean field: the expectation value $\langle\phi\rangle$ of the fundamental microscopic variable. Its equation of motion $\delta\Gamma/\delta\phi_{\text{cl}} = 0$ is the quantum-corrected classical field equation (e.g., the Klein–Gordon equation with loop corrections, or the Ginzburg–Landau equation for the order parameter). This saddle describes the *average behavior of the microscopic degrees of freedom*—the average position of each bird, the average behavior of each person.

The saddle of $S_{\text{eff}}[\sigma]$ is the most probable value of the collective/auxiliary field: $\bar{\sigma}$ satisfying $\delta S_{\text{eff}}/\delta\sigma = 0$. This $\bar{\sigma}$ is not an expectation value of a fundamental field; it is the mode of the marginal distribution over a *composite* variable introduced by the HST. The gap equation determines $\bar{\sigma}$ self-consistently in terms of the microscopic propagator. This saddle describes the *macroscopic collective pattern*—the flock shape, the societal norm, the pairing gap, the screened potential.

The distinction matters because fluctuations around the two saddles describe different physics:

	Schwinger saddle	**HST saddle**
Saddle field	$\phi_{\text{cl}} = \langle\phi\rangle$	$\bar{\sigma} = \text{mode}(p(\sigma))$
Equation	$\delta\Gamma/\delta\phi_{\text{cl}} = 0$	$\delta S_{\text{eff}}/\delta\sigma = 0$ (gap eq.)
Describes	Average microscopic behavior	Collective pattern / order parameter
Fluctuations	Quasiparticle excitations	Collective-mode excitations
Hessian gives	G_ϕ^{-1} (microscopic precision)	$K = D_0^{-1} + \Pi$ (collective precision)

In the person/society analogy: the saddle of Γ tells you what the average person does (their daily routine, consumption patterns, social behavior). The saddle of S_{eff} tells you what the society looks like (its institutions, norms, laws). Fluctuations around the Γ saddle are individual deviations from average behavior—these are "quasiparticle" excitations, and they cost at least Δ (the gap, the conformity cost). Fluctuations around the S_{eff} saddle are shifts in the collective norm—these are collective modes, and some (the Goldstone modes) cost nothing.

This is the mathematical origin of the book's central result: the same physical system can have gapped quasiparticle excitations (large Hessian eigenvalues of Γ) and gapless collective modes (zero Hessian eigenvalues of S_{eff}) simultaneously. The two saddles are different objects, their Hessians are different operators, and there is no contradiction.

To summarize in the sharpest possible terms:

The **gap equation** is about society: it determines the self-consistent collective norm $\bar{\sigma}$ that sustains itself. The **gap** Δ is about the individual: it is the minimum cost a person (quasiparticle) must pay to break free from the norm.
But the two are not independent. The norm $\bar{\sigma}$ determines the gap Δ (through the modified precision $M(\bar{\sigma}) = G_0^{-1} + V(\bar{\sigma})$), and the gap feeds back into the norm (through the prop-

agator $G_{\bar{\sigma}}$ in the gap equation). Society shapes the cost of nonconformity, and the cost of nonconformity sustains society. This self-consistency—the fact that the collective pattern and the individual cost are determined simultaneously by the same equation—is the essence of mean-field theory.

Chapter 14

The Gap as a Spectral or Covariance Scale

14.1 How the Saddle Modifies the Microscopic Precision

Recall from Parts II and III that the effective action typically arises from integrating out microscopic fields χ whose precision depends on the collective field σ:

$$M(\sigma) = A_0 + V(\sigma),$$

where A_0 is the bare (reference) precision and $V(\sigma)$ encodes the coupling to σ. At the saddle $\bar{\sigma}$, the microscopic precision becomes

$$M(\bar{\sigma}) = A_0 + V(\bar{\sigma}).$$

The corresponding covariance—the propagator for the microscopic field—is

$$G_{\bar{\sigma}} = M(\bar{\sigma})^{-1} = (A_0 + V(\bar{\sigma}))^{-1}. \tag{14.1}$$

This is the **saddle-dressed** or **mean-field propagator**: the covariance of the microscopic field, computed at the saddle-point value of the collective field. It differs from the bare propagator $G_0 = A_0^{-1}$ because the saddle has modified the precision.

14.2 The Gap in Fourier Space

To understand the physical meaning of this modification, let us work in Fourier space for a translation-invariant system. Suppose the bare precision in Fourier space is $\tilde{A}_0(k) = k^2 + m_0^2$, so the bare propagator is $\tilde{G}_0(k) = 1/(k^2 + m_0^2)$, with correlation length $\xi_0 = 1/m_0$.

Now suppose the saddle shifts the precision by a constant Δ^2:

$$\tilde{M}(\bar{\sigma}, k) = k^2 + m_0^2 + \Delta^2.$$

The saddle-dressed propagator is

$$\tilde{G}_{\bar{\sigma}}(k) = \frac{1}{k^2 + m_0^2 + \Delta^2} = \frac{1}{k^2 + m^2}, \tag{14.2}$$

where $m^2 = m_0^2 + \Delta^2$ is the total mass squared. The parameter Δ is the **gap**—a mass scale generated by the saddle.

14.3 What "Gap" Means

The term "gap" has a precise spectral meaning. The propagator $\tilde{G}(k) = 1/(k^2 + m^2)$ has a pole (in the analytically continued version) at $k^2 = -m^2$, or equivalently, the spectrum of the precision operator $-\nabla^2 + m^2$ is bounded below by $m^2 > 0$. There is a gap of size m^2 between zero and the lowest eigenvalue.

In real (Euclidean) space, the gap manifests as exponential decay of the covariance:

$$G_{\bar{\sigma}}(x) \sim \frac{e^{-m|x|}}{|x|^{(d-1)/2}} \quad \text{as } |x| \to \infty,$$

where d is the spatial dimension. The correlation length is $\xi = 1/m$. A nonzero gap Δ shortens the correlation length relative to the bare value $1/m_0$.

Physicists use many names for the same concept, depending on the context: *mass* or *mass gap* (relativistic field theory), *quasiparticle gap* (condensed matter), *pairing gap* (superconductivity), *spectral gap* (transfer matrices), *inverse correlation length* (statistical mechanics), *dynamically generated mass* (when absent from the bare action). Despite the variety of names, the mathematical content is the same: the saddle has modified the precision operator so that its spectrum is bounded below by a positive number, and the corresponding covariance decays exponentially.

Chapter 15

The Gap Equation from the First Derivative of the Effective Action

15.1 Deriving the Gap Equation

We now derive the equation that determines the gap. This is the saddle-point condition (13.1) applied to the specific form of the effective action from Chapter 6.

Consider the bosonic effective action:

$$S_{\text{eff}}[\sigma] = S_0[\sigma] + \frac{1}{2}\operatorname{Tr}\log M(\sigma),$$

where $S_0[\sigma]$ is the bare action for σ and $M(\sigma) = A_0 + V(\sigma)$ is the σ-dependent precision of the integrated-out field. The saddle condition requires

$$\left.\frac{\delta S_{\text{eff}}}{\delta\sigma(x)}\right|_{\bar\sigma} = 0.$$

We compute the two terms separately.

The first term gives $\frac{\delta S_0}{\delta\sigma(x)}|_{\bar\sigma}$, which depends on the specific form of S_0. For the common choice $S_0[\sigma] = \frac{1}{2g}\int \sigma(x)^2\,\mathrm{d}x$, this is simply $\bar\sigma(x)/g$.

The second term requires differentiating $\frac{1}{2}\operatorname{Tr}\log M(\sigma)$ with respect to $\sigma(x)$. By Jacobi's formula (11.3):

$$\frac{\delta}{\delta\sigma(x)}\frac{1}{2}\operatorname{Tr}\log M(\sigma) = \frac{1}{2}\operatorname{Tr}\left(M(\sigma)^{-1}\frac{\delta M(\sigma)}{\delta\sigma(x)}\right).$$

For the common case $M(\sigma) = A_0 + \sigma\Gamma$ (linear dependence on σ), we have $\frac{\delta M}{\delta\sigma(x)} = \Gamma\cdot\delta(x)$, and the trace reduces to

$$\frac{1}{2}\operatorname{tr}[G_\sigma(x,x)\Gamma]\,,$$

where $G_\sigma = M(\sigma)^{-1}$ is the propagator and tr denotes a trace over internal indices. The notation $G_\sigma(x,x)$ means the kernel evaluated at coincident points. In momentum space, $G_\sigma(x,x) = \int_k G_\sigma(k)$.

Putting both terms together, the saddle equation for $S_0[\sigma] = \frac{1}{2g}\int\sigma^2$ is

$$\boxed{\frac{\bar\sigma(x)}{g} = -\frac{1}{2}\operatorname{tr}[G_{\bar\sigma}(x,x)\Gamma]\,.} \tag{15.1}$$

For a uniform (translation-invariant) saddle $\bar{\sigma}(x) = \bar{\sigma}$, the propagator $G_{\bar{\sigma}}$ is also translation-invariant, and $G_{\bar{\sigma}}(x, x) = \int_k G_{\bar{\sigma}}(k)$. The gap equation becomes

$$\boxed{\frac{\bar{\sigma}}{g} = -\frac{1}{2} \int_k \mathrm{tr}[\Gamma\, G_{\bar{\sigma}}(k)]\,.} \tag{15.2}$$

15.2 Anatomy of the Gap Equation

The gap equation (15.2) is a self-consistency relation. The left-hand side involves the saddle value $\bar{\sigma}$. The right-hand side involves the propagator $G_{\bar{\sigma}} = (A_0 + \bar{\sigma}\Gamma)^{-1}$, which itself depends on $\bar{\sigma}$. So the equation is implicit: $\bar{\sigma}$ appears on both sides.

In statistical language, the gap equation says: the saddle value of the latent field equals the coupling constant g times the conditional expectation of the composite observable $O(\phi) = \mathrm{tr}[\Gamma\phi\phi^\top]$, evaluated at the saddle-dressed covariance.

The left-hand side, $\bar{\sigma}/g$, is the derivative of the bare action $S_0 = \bar{\sigma}^2/(2g)$. It represents a restoring force that pushes σ back toward zero. The right-hand side is the "pull" from the microscopic degrees of freedom—the tendency of the fluctuations to sustain a nonzero value of σ.

If the right-hand side is strong enough to overcome the restoring force, a nonzero solution $\bar{\sigma} \neq 0$ exists, and the system develops a gap. If not, the only solution is $\bar{\sigma} = 0$, and no gap is generated. The transition between these two regimes—the appearance or disappearance of a nonzero saddle—is a **phase transition** in the physical interpretation.

15.3 A One-Dimensional Example

To see the gap equation in its simplest form, consider n scalar fields χ_i with precision $M(\sigma) = (m_0^2 + \sigma)\mathbf{1}_{n\times n}$, where σ is a single real parameter and $\Gamma = \mathbf{1}$. The bare action is $S_0(\sigma) = \sigma^2/(2g)$. The propagator at the saddle is $G_{\bar{\sigma}} = (m_0^2 + \bar{\sigma})^{-1}\mathbf{1}$, and

$$\mathrm{Tr}(G_{\bar{\sigma}}\Gamma) = \frac{n}{m_0^2 + \bar{\sigma}}.$$

The gap equation (15.2) becomes

$$\frac{\bar{\sigma}}{g} = \frac{n}{2(m_0^2 + \bar{\sigma})}.$$

Rearranging: $2\bar{\sigma}(m_0^2 + \bar{\sigma}) = ng$, or $2\bar{\sigma}^2 + 2m_0^2\bar{\sigma} - ng = 0$. The quadratic formula gives

$$\bar{\sigma} = -\frac{m_0^2}{2} + \frac{1}{2}\sqrt{m_0^4 + 2ng}.$$

For small coupling g, $\bar{\sigma} \approx ng/(2m_0^2)$: the gap is proportional to the number of fields times the coupling, divided by the bare mass squared. For large n, the gap grows as $\sqrt{ng}$, confirming that the saddle becomes increasingly important. This simple example illustrates the general pattern: the gap is determined by a self-consistency relation between the saddle value and the propagator evaluated at that saddle.

Chapter 16

The Gap Equation in Terms of Dressed Propagators

16.1 Three Levels of "Dressing"

Before writing the gap equation more explicitly, we must clarify a source of persistent confusion in the physics literature: the word "dressed," applied to a propagator, can mean at least three different things.

The bare propagator is the covariance of the Gaussian reference distribution:

$$G_0 = A_0^{-1}.$$

This is the propagator with no interactions, no saddle, no self-energy corrections of any kind.

The saddle-dressed (mean-field) propagator is the covariance evaluated at the saddle value of the collective field:

$$G_{\bar{\sigma}} = (A_0 + V(\bar{\sigma}))^{-1} = \left(G_0^{-1} + V(\bar{\sigma})\right)^{-1}.$$

This includes the effect of the saddle—the mean-field background—but not the effect of fluctuations around the saddle. In the gap equation (15.2), the propagator that appears on the right-hand side is saddle-dressed.

The fully dressed propagator includes both the saddle and fluctuation corrections:

$$G = \left(G_0^{-1} + V(\bar{\sigma}) + \Sigma_{\text{fluc}}\right)^{-1},$$

where Σ_{fluc} represents additional self-energy corrections from fluctuations of σ around the saddle. Computing Σ_{fluc} requires going beyond the saddle-point approximation.

16.2 The Self-Consistent Gap Equation

A more ambitious equation replaces the saddle-dressed propagator by the fully dressed one:

$$\frac{\Delta}{g} = \int_k \text{tr}[\Gamma\, G(k)]\,,$$

where G includes fluctuation corrections. The logic is appealing: if the gap equation is a self-consistency condition, and if the propagator that mediates the self-consistency should include all the physics, then we should use the fully dressed G.

However, upgrading the gap equation in this way is not automatically an improvement. If the fluctuation self-energy Σ_{fluc} is computed with one set of approximations while the gap equation uses a different level, the resulting system may violate exact identities—conservation laws, Ward identities, sum rules—that the exact theory satisfies. We will return to this in Chapter 32, where we discuss conserving approximations.

For now, the important takeaway is this: whenever you encounter a "gap equation" in the literature, check which propagator is being used on the right-hand side—bare, saddle-dressed, or fully dressed—and whether the approximation scheme is internally consistent.

What was exact, what was approximate, and what was conventional?

Exact. The effective action is exact when the integrated-out field is exactly Gaussian (as in the HST framework). The stationarity condition (13.1) is the exact definition of a critical point.

Approximate. The saddle-point approximation itself—replacing the full integral over σ by the contribution from $\bar{\sigma}$—is an approximation, valid when fluctuations of σ are small.

Conventional. The terms "gap equation," "mean-field equation," and "self-consistency equation" are used interchangeably. The choice of sign conventions varies between subfields.

Part V

Hessians, Polarization, Self-Energy, and Dressed Propagators

This part develops the Gaussian fluctuation theory around the saddle point, deriving the collective Hessian, the polarization bubble, the self-energy, the Dyson equation, and the RPA in a unified framework. Every result is first stated as a matrix identity, then interpreted statistically, and only then given its physics name.

Chapter 17

Gaussian Fluctuations Around the Saddle

17.1 Expanding Around the Saddle

We have found the saddle $\bar{\sigma}$ in Part IV. Now we study fluctuations around it. Write

$$\sigma(x) = \bar{\sigma}(x) + \eta(x),$$

where η is the fluctuation. Expand the effective action to second order in η:

$$S_{\text{eff}}[\bar{\sigma} + \eta] = S_{\text{eff}}[\bar{\sigma}] + \underbrace{\left.\frac{\delta S_{\text{eff}}}{\delta \sigma}\right|_{\bar{\sigma}} \cdot \eta}_{=0 \text{ (saddle condition)}} + \frac{1}{2}\eta^{\top} K \eta + O(\eta^3),$$

where

$$K(x, y) = \left.\frac{\delta^2 S_{\text{eff}}}{\delta\sigma(x)\,\delta\sigma(y)}\right|_{\bar{\sigma}} \tag{17.1}$$

is the **Hessian** of the effective action at the saddle. The linear term vanishes because $\bar{\sigma}$ is a critical point.

Within the Gaussian (quadratic) approximation, the fluctuation η is a Gaussian random variable with precision K. Its covariance is therefore

$$\boxed{D = K^{-1}.} \tag{17.2}$$

This is the Gaussian covariance of the collective-field fluctuations around the saddle. Physicists call D the **dressed propagator of the collective field**.

Let us be very precise about the logical status of this statement. The effective action $S_{\text{eff}}[\sigma]$ defines a probability distribution over σ. At the saddle, we approximate this distribution by a Gaussian with precision K. The inverse of that precision is the covariance D. This is a Laplace approximation—the same technique we used in Chapter 8, but now applied to the collective field rather than to the original microscopic field.

17.2 What K and D Describe

The Hessian K and the covariance $D = K^{-1}$ describe different aspects of the collective-field distribution.

$K(x, y)$ is the **precision** of the collective field. It tells us how costly it is (in units of action, or negative log-probability) for $\eta(x)$ and $\eta(y)$ to deviate from zero simultaneously. Large eigenvalues of K correspond to "stiff" directions—fluctuations in those directions are strongly suppressed.

$D(x, y) = K^{-1}(x, y)$ is the **covariance** of the collective field. It tells us how much $\eta(x)$ and $\eta(y)$ actually fluctuate together. Small eigenvalues of K—corresponding to soft directions—produce large eigenvalues of D, meaning that fluctuations in those directions are large and long-ranged.

A zero eigenvalue of K (a flat direction in S_{eff}) produces a divergent eigenvalue of D. In physical language: a direction along which the effective action has no curvature is a direction in which fluctuations are completely unconstrained. This is the hallmark of a **Goldstone mode** or a **soft mode**, which we will study in Part VI.

Chapter 18

The Polarization Bubble as a Hessian

18.1 Computing the Hessian

We now compute the Hessian K for the effective action

$$S_{\text{eff}}[\sigma] = S_0[\sigma] + \frac{1}{2}\,\mathrm{Tr}\log M(\sigma),$$

with $M(\sigma) = A_0 + \sigma\Gamma$ (linear dependence on σ). The Hessian has two contributions:

$$K = \left.\frac{\delta^2 S_0}{\delta\sigma^2}\right|_{\bar\sigma} + \frac{1}{2}\frac{\delta^2}{\delta\sigma^2}\,\mathrm{Tr}\log M(\sigma)\bigg|_{\bar\sigma}.$$

The first term is the Hessian of the bare action. For $S_0 = \frac{1}{2g}\int\sigma^2$, this is simply $\frac{1}{g}\delta(x-y)$, or in Fourier space, a constant $1/g$ for all momenta. This is the **bare inverse propagator** of the collective field:

$$D_0^{-1} = \frac{1}{g}.$$

The second term involves the second derivative of $\mathrm{Tr}\log M$ with respect to σ, which we already computed in Chapter 9 (equation 11.8). For $M = A_0 + \sigma\Gamma$ with $\delta^2 M = 0$:

$$\frac{1}{2}\frac{\delta^2}{\delta\sigma(x)\,\delta\sigma(y)}\,\mathrm{Tr}\log M = -\frac{1}{2}\,\mathrm{tr}[G_{\bar\sigma}(x,y)\Gamma\,G_{\bar\sigma}(y,x)\Gamma]\,,$$

where $G_{\bar\sigma} = M(\bar\sigma)^{-1}$ is the saddle-dressed propagator. Defining the **polarization** (with sign conventions absorbed):

$$\Pi(x,y) = \mathrm{tr}[G_{\bar\sigma}(x,y)\Gamma\,G_{\bar\sigma}(y,x)\Gamma]\,, \tag{18.1}$$

the Hessian becomes

$$\boxed{K = D_0^{-1} + \Pi.} \tag{18.2}$$

18.2 The Polarization Bubble in Momentum Space

For a translation-invariant system, the polarization depends only on the difference $x - y$, and its Fourier transform is

$$\boxed{\Pi(q) = \int_k \mathrm{tr}[G_{\bar\sigma}(k+q)\,\Gamma\,G_{\bar\sigma}(k)\,\Gamma]\,.} \tag{18.3}$$

This is the **polarization bubble** in momentum space: two saddle-dressed propagators, carrying momenta $k+q$ and k, connected at two vertices Γ, with the internal momentum k integrated over.

18.3 Three Equivalent Interpretations

We have now arrived at the polarization bubble from three different starting points, each giving the same algebraic expression:

Interpretation 1: Second derivative of a log-determinant. The bubble is $-\frac{1}{2}\frac{\delta^2}{\delta\sigma^2}\operatorname{Tr}\log M(\sigma)|_{\bar{\sigma}}$—a Hessian contribution to the effective action.

Interpretation 2: Gaussian covariance of a composite observable. The bubble is $\mathrm{Cov}_0(\phi^2(x), \phi^2(y)) = 2C(x,y)^2$ (from Chapter 3)—the Gaussian covariance of the composite observable that couples to the collective field.

Interpretation 3: Linear response or susceptibility. The bubble gives the leading-order response of $\langle O \rangle$ to an external source coupled to O (from the fluctuation-response identity).

These three interpretations are not three independent results. They are three viewpoints on a single mathematical object. The equivalence arises because the Hubbard–Stratonovich construction couples the collective field σ to a composite observable $O(\phi)$, and differentiating the effective action with respect to σ probes how O responds to changes in the background—which is exactly its covariance.

Chapter 19

Polarization as Response

19.1 The Response Function Perspective

There is a complementary way to derive the polarization that does not go through the Hubbard–Stratonovich transformation. Consider a probability distribution $p(\phi) \propto e^{-S[\phi]}$ and a composite observable $O[\phi]$. Add an external source J coupled to O:

$$S_J[\phi] = S[\phi] - J \cdot O[\phi].$$

The generating functional is $W[J] = \log \int \mathcal{D}\phi \, e^{-S[\phi]+J\cdot O[\phi]}$, and

$$\frac{\delta W}{\delta J(x)} = \mathbb{E}_J[O(x)], \qquad \frac{\delta^2 W}{\delta J(x)\, \delta J(y)} = \mathrm{Cov}_J(O(x), O(y))\,.$$

Evaluated at $J = 0$, the second derivative gives the equilibrium covariance of O. It is also the **linear response function**: if we perturb J slightly, the change in $\langle O(x)\rangle$ to first order is

$$\delta\langle O(x)\rangle = \int \mathrm{Cov}(O(x), O(y))\, \delta J(y)\, \mathrm{d}y.$$

This identity—fluctuation equals response—holds exactly, for any distribution of the form e^{-S}.

19.2 The Physics Vocabulary

Different choices of O give the response function different names in physics. When $O = \rho$ (charge density), the covariance $\mathrm{Cov}(\rho, \rho)$ is the **density response** or **polarization** Π. When $O = m$ (magnetization density), the covariance is the **magnetic susceptibility** χ. When $O = j$ (current density), the covariance is the **current-current response**, related to the **conductivity**. When $O = \phi$ (the field itself), the covariance is simply the two-point function $C = \mathrm{Cov}(\phi, \phi)$, the propagator.

In each case, the mathematical content is the same: we are computing the covariance of an observable. The different names reflect the different physical quantities that O represents, not different mathematical structures.

Chapter 20

Self-Energy as a Precision Correction

20.1 Definition

We now introduce one of the most important concepts in field theory—the self-energy—in its natural mathematical setting. The definition is simple:

The self-energy is the difference between the exact precision and the reference precision.

Let C be the exact covariance (propagator) of some field, and let C_0 be the covariance of a reference Gaussian distribution. Define

$$\boxed{\Sigma = C^{-1} - C_0^{-1}.} \tag{20.1}$$

That is all. The self-energy Σ is a correction to the inverse covariance—a precision correction. It encodes everything that distinguishes the exact two-point statistics from the reference statistics.

Rearranging:

$$C^{-1} = C_0^{-1} + \Sigma,$$

so the exact precision equals the reference precision plus the self-energy. Inverting:

$$\boxed{C = (C_0^{-1} + \Sigma)^{-1}.} \tag{20.2}$$

This is the **Dyson equation**.

20.2 What the Self-Energy Tells Us

The precision C^{-1} describes the cost of fluctuations in the exact distribution. The reference precision C_0^{-1} describes the cost in the Gaussian reference. Their difference Σ is the additional cost (or reduced cost, if Σ has negative eigenvalues) arising from everything beyond the reference model—interactions, mean-field backgrounds, fluctuation corrections.

In Fourier space for a translation-invariant system:

$$\tilde{\Sigma}(k) = \tilde{C}(k)^{-1} - \tilde{C}_0(k)^{-1}.$$

If $\tilde{C}_0(k) = 1/(k^2 + m_0^2)$ and $\tilde{C}(k) = 1/(k^2 + m_0^2 + \Sigma(k))$, then $\Sigma(k)$ shifts the pole of the propagator: the effective mass becomes $m^2 = m_0^2 + \Sigma(0)$ at zero momentum, and the momentum dependence of $\Sigma(k)$ modifies the dispersion relation and introduces damping.

20.3 An Important Distinction

Within the Gaussian approximation for the collective field, the self-energy of the collective field equals the polarization bubble: $\Sigma_D = \Pi$. But this identification holds only at the RPA level. In general, the exact self-energy includes corrections beyond the bubble, coming from higher-order terms in the Taylor expansion of S_{eff} (vertex corrections, self-energy insertions, etc.).

Furthermore, the bubble itself can be computed with different propagators: bare (G_0), saddle-dressed ($G_{\bar{\sigma}}$), or fully dressed (G). These choices give different numerical values and correspond to different levels of approximation.

Self-energy is the exact correction to the inverse covariance, defined non-perturbatively by $\Sigma = C^{-1} - C_0^{-1}$. The polarization bubble is a specific algebraic expression that equals the self-energy in certain approximations (Gaussian fluctuation theory, RPA) but not in general.

Chapter 21

The Dyson Equation and the Neumann Expansion

21.1 The Dyson Equation as an Algebraic Identity

The Dyson equation

$$C = (C_0^{-1} + \Sigma)^{-1}$$

is an immediate consequence of the definition $\Sigma = C^{-1} - C_0^{-1}$. It is an algebraic identity, not a physical postulate and not a resummation. Let us state this with emphasis, because the traditional physics pedagogy sometimes presents the Dyson equation as a profound result derived through an infinite resummation of diagrams:

> The Dyson equation is not conceptually prior to the self-energy. Once the self-energy is defined as a precision correction, the Dyson equation is an algebraic identity, and the infinite series that follows from it is simply the Neumann expansion of a matrix inverse.

21.2 The Neumann Expansion

The Neumann expansion (the matrix analogue of the geometric series) is obtained by factoring out C_0^{-1}:

$$C = (C_0^{-1}(I + C_0\Sigma))^{-1} = (I + C_0\Sigma)^{-1}C_0.$$

If $\|C_0\Sigma\| < 1$, we expand $(I + X)^{-1} = I - X + X^2 - X^3 + \cdots$:

$$C = C_0 - C_0\Sigma C_0 + C_0\Sigma C_0\Sigma C_0 - C_0\Sigma C_0\Sigma C_0\Sigma C_0 + \cdots \tag{21.1}$$

This is the **Neumann series** for the corrected covariance. Each term has n insertions of Σ, separated by bare propagators C_0.

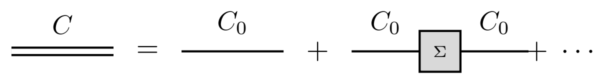
Physicists call this the **Dyson series** or **Dyson resummation**, and they draw it diagrammatically as a chain of self-energy blobs connected by bare propagator lines. From our perspective, the closed form (20.2) is the definition, and the series (21.1) is its expansion.

Note the sign of the first correction: $C \approx C_0 - C_0\Sigma C_0 + \cdots$. Let us verify that this is consistent with $C^{-1} = C_0^{-1} + \Sigma$ to first order. Compute:

$$(C_0^{-1} + \Sigma)(C_0 - C_0\Sigma C_0) = I - \Sigma C_0 + \Sigma C_0 - \Sigma C_0\Sigma C_0 = I - \Sigma C_0\Sigma C_0.$$

The remaining term is second order in Σ, confirming the expansion is correct.

Diagrammatically, the Dyson equation can be drawn as:



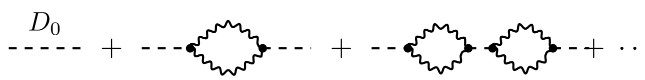
The double line represents the exact (dressed) propagator C. Single lines are bare propagators C_0. The shaded box is the self-energy Σ. The series continues with two, three, and more insertions.

For the collective field with $D_0^{-1} + \Pi$ as the Hessian, and for a contact bare interaction $D_0 = g$ (a constant):

$$D(q) = \frac{g}{1 + g\Pi(q)} = \frac{1}{g^{-1} + \Pi(q)} = \frac{1}{D_0^{-1} + \Pi(q)}.$$

The bubble chain $D = g - g\Pi g + g\Pi g\Pi g - \cdots$ is its geometric-series representation.

Diagrammatically, the RPA bubble chain is:



Dashed lines represent the bare collective propagator D_0. Each bubble (a pair of wavy propagator lines) is one insertion of the polarization Π. The full sum gives the dressed collective propagator $D = (D_0^{-1} + \Pi)^{-1}$.

Chapter 22

RPA as Gaussian Inference

22.1 The Full RPA Procedure

We are now in a position to present the Random Phase Approximation (RPA) as a coherent approximation scheme. The procedure consists of five steps, each of which we have already developed:

Step 1: Introduce a collective field σ (HST, exact).

Step 2: Integrate out the microscopic field (exact if quadratic), producing $S_{\text{eff}}[\sigma] = S_0[\sigma] \pm \operatorname{Tr}\log M(\sigma) + \text{const}$.

Step 3: Find the saddle $\bar{\sigma}$ by solving $\delta S_{\text{eff}}/\delta\sigma = 0$ (approximate: saddle-point). This gives the gap equation.

Step 4: Compute the Hessian at the saddle: $K = D_0^{-1} + \Pi$ (approximate: Gaussian truncation). The polarization Π is a bubble.

Step 5: Invert to get the dressed collective propagator: $D = K^{-1}$ (RPA).

The result:

$$\boxed{\text{RPA} = \text{Gaussian (Laplace) approximation for collective-field fluctuations.}} \tag{22.1}$$

The "bubble chain" is simply the Neumann expansion of K^{-1}.

22.2 RPA as a Variational Approximation

The RPA can also be understood from a variational perspective. The **Laplace approximation**—matching a Gaussian to the mode and Hessian of S_{eff}—is a specific approximation scheme, and it is what the RPA amounts to. However, it is important to distinguish this from the **optimal Gaussian variational approximation**, which minimizes $D_{\text{KL}}(q\|p)$ over all Gaussian q.

The optimal reverse-KL Gaussian satisfies the stationarity conditions

$$\mathbb{E}_q[\nabla_\sigma S_{\text{eff}}(\sigma)] = 0, \qquad C^{-1} = \mathbb{E}_q[\nabla_\sigma^2 S_{\text{eff}}(\sigma)],$$

where C is the variational covariance. The Laplace approximation replaces these expectations by evaluations at the mode: $\nabla S_{\text{eff}}(\bar{\sigma}) = 0$ and $C^{-1} = \nabla^2 S_{\text{eff}}(\bar{\sigma})$. The two coincide when the distribution is sharply peaked (so that the expectations are dominated by the mode), which is the same regime—large N, many degrees of freedom—where the saddle-point approximation is justified.

When the distribution of σ is not sharply peaked, the Laplace and optimal variational Gaussian differ: the variational mean shifts away from the mode, and the variational precision averages the Hessian over the distribution rather than evaluating it at a point. In practice, the Laplace/RPA approximation is simpler and suffices whenever the saddle is justified; the variational Gaussian is more robust but requires solving a self-consistent equation.

From the machine-learning perspective, the RPA is a form of **Laplace inference**: it approximates a complicated posterior by a Gaussian centered at the MAP estimate with precision equal to the Hessian. This is asymptotically optimal (as $N \to \infty$) but not globally optimal for finite N.

22.3 The RPA Collective-Mode Spectrum

The poles of D (zeros of $\det K$) are the collective modes:

$$\det K(q,\omega) = 0 \qquad \Longleftrightarrow \qquad D_0^{-1}(q) + \Pi(q,\omega) = 0.$$

This condition determines the dispersion relation $\omega(q)$ of the collective mode. The physical nature of the mode depends on what σ represents: sound, plasma oscillation, spin wave, phase oscillation, etc.

22.4 Screening

When $D_0(q) = V(q)$ is a bare interaction potential (e.g., the Coulomb potential), the dressed propagator $D(q) = V(q)/(1 + V(q)\Pi(q))$ is the **screened interaction**. The denominator $\varepsilon(q) = 1 + V(q)\Pi(q)$ is the **dielectric function**. The bare long-range interaction is reduced by $\varepsilon(q)$, converting it into a short-range effective interaction. This is the physical content of screening: microscopic degrees of freedom rearrange in response to an external charge, partially canceling its potential.

But all of this physical storytelling is an interpretation of the mathematical result $D = (D_0^{-1} + \Pi)^{-1}$. The mathematics is an inverse of a corrected precision operator. The physical story explains why this mathematical structure is relevant for specific systems.

What was exact, what was approximate, and what was conventional?

Exact. The HST identity. The Gaussian integration producing $\operatorname{Tr}\log$. The definition $\Sigma = C^{-1} - C_0^{-1}$. The Dyson equation (20.2). The Neumann series (21.1) within its convergence radius.

Approximate. The saddle-point approximation (Step 3). The Gaussian truncation at the Hessian (Steps 4–5). The computation of Π with saddle-dressed rather than fully dressed propagators. The neglect of vertex corrections.

Conventional. The term “RPA” for the Gaussian approximation. “Dyson resummation” for the Neumann series. “Bubble chain” for its diagrammatic representation. The various sign conventions for the self-energy.

Part VI

Gaps, Zero Modes, and Hydrodynamic Behavior

The previous two parts developed a systematic framework: introduce a collective field, integrate out microscopic variables, find a saddle, compute the Hessian, and invert it to obtain a collective propagator. The saddle modifies the microscopic precision and may generate a gap. The Hessian determines the collective-field covariance.

This part addresses a question that is central to modern physics: can the same saddle simultaneously produce a gap in one covariance and a gapless mode in another? The answer is yes, and the mathematical reason is simple—the microscopic covariance and the collective covariance are different operators acting on different variables.

Chapter 23

The Same Saddle Can Produce a Gap and a Gapless Mode

23.1 Two Covariance Operators from One Saddle

The saddle $\bar{\sigma}$ modifies two distinct objects. The **microscopic covariance** is the propagator of ϕ, evaluated at the saddle:

$$G_{\bar{\sigma}} = \left(G_0^{-1} + V(\bar{\sigma})\right)^{-1}.$$

If $V(\bar{\sigma})$ shifts the precision by a positive amount at all momenta, then $G_{\bar{\sigma}}$ is gapped.

The **collective covariance** is the inverse Hessian of S_{eff} at the saddle:

$$D = K^{-1} = \left(D_0^{-1} + \Pi\right)^{-1}.$$

Whether D is gapped or gapless depends on the eigenvalue spectrum of K. If K has a zero eigenvalue, D is gapless.

These are different operators. $G_{\bar{\sigma}}$ acts on the space of microscopic field configurations. D acts on the space of collective field configurations. There is no mathematical reason for their spectral properties to be the same.

A symmetry-breaking saddle simultaneously gaps the microscopic covariance and produces a zero-Hessian direction in the collective covariance. These are not contradictory because they refer to different operators on different variables.

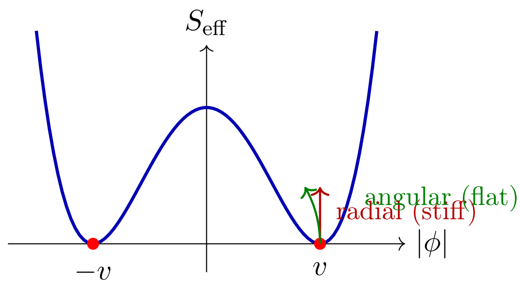
23.2 A Finite-Dimensional Analogy

Consider a two-dimensional density $p(x, y) \propto \exp(-\frac{\lambda}{4}(x^2 + y^2 - v^2)^2)$ for some $v > 0$. This density is rotationally symmetric. Its mode lies on the circle $x^2 + y^2 = v^2$.

Pick any mode, say $(\bar{x}, \bar{y}) = (v, 0)$. In radial-angular coordinates $x = (v + \rho)\cos\theta$, $y = (v + \rho)\sin\theta$, the density near the mode becomes (to second order):

$$p \propto \exp(-\lambda v^2 \rho^2).$$

The radial direction ρ has nonzero Hessian eigenvalue $2\lambda v^2$—it is **stiff** (gapped). The angular direction θ has zero Hessian eigenvalue—it is **flat** (gapless).



This two-dimensional example captures the essential structure. Replace ρ with the amplitude fluctuation of a complex order parameter and θ with its phase, and you have the Goldstone phenomenon in field theory.

Chapter 24

Symmetry Orbits and Goldstone Zero Modes

24.1 The Goldstone Theorem

Suppose $S_{\text{eff}}[U(\alpha)\sigma] = S_{\text{eff}}[\sigma]$ for all α. If $\bar{\sigma}$ is a saddle, define the tangent vector to the symmetry orbit:

$$v(x) = \frac{d}{d\alpha}\bigg|_{\alpha=0} [U(\alpha)\bar{\sigma}](x).$$

Since $f(\alpha) = S_{\text{eff}}[U(\alpha)\bar{\sigma}]$ is constant (by symmetry), $f''(0) = 0$. Computing $f''(0)$ via the chain rule:

$$f''(0) = v^\top K v = 0$$

(the terms involving $\frac{\delta S_{\text{eff}}}{\delta \sigma}|_{\bar{\sigma}}$ drop out because of the saddle condition). Since K is positive semi-definite at a stable saddle, $v^\top K v = 0$ implies

$$\boxed{Kv = 0.} \tag{24.1}$$

The tangent to the symmetry orbit is a **null eigenvector** of the Hessian. This is the **Goldstone theorem**: a purely algebraic consequence of (i) a continuous symmetry of S_{eff} and (ii) a saddle that breaks it.

In relativistic systems, the number of Goldstone modes equals $\dim G - \dim H$, where G is the symmetry group and H is the stabilizer of the saddle. For $O(N) \to O(N-1)$, this gives $N-1$ Goldstone modes. For $U(1)$ completely broken, this gives one.

In nonrelativistic systems, the counting is different and requires more care. Define the **commutator-density matrix**

$$\rho_{ab} = -\frac{\mathrm{i}}{V}\langle [Q_a, Q_b] \rangle,$$

where Q_a are the broken generators and V is the volume. The Watanabe–Murayama counting rule gives the number of independent Nambu–Goldstone modes as

$$N_{\text{NG}} = N_{\text{BS}} - \frac{1}{2}\operatorname{rank}\rho, \tag{24.2}$$

where $N_{\text{BS}} = \dim G - \dim H$ is the number of broken symmetry generators. The modes fall into two classes:

Type-A (unpaired): each broken generator gives one mode with *linear* dispersion $\omega \propto q$. These arise when $\rho_{ab} = 0$ for the corresponding generators. Example: the two transverse spin-wave modes of an *antiferromagnet*, where the staggered-magnetization order parameter commutes with the total spin (which vanishes in the ground state).

Type-B (paired): two broken generators pair up to give one mode with *quadratic* dispersion $\omega \propto q^2$. These arise when $\rho_{ab} \neq 0$, indicating that the two generators do not commute in the ground state. Example: the single magnon mode of a *ferromagnet*, where two broken spin-rotation generators (S_x and S_y) satisfy $\langle [S_x, S_y] \rangle = \mathrm{i}\langle S_z \rangle \neq 0$, and pair into one mode with $\omega \propto q^2$.

The Hessian analysis of this book gives the correct counting automatically: the number of zero eigenvalues of K at $q = 0$ is N_{NG}, not N_{BS}. The distinction between linear and quadratic dispersion arises from whether the effective action contains a first-order or second-order time derivative, which is determined by the symmetry structure through ρ_{ab}.

Chapter 25

The Hydrodynamic Field from a Slowly Varying Saddle Coordinate

25.1 Promoting the Symmetry Coordinate to a Field

The Goldstone theorem tells us that the Hessian has a zero eigenvalue at zero momentum. But in a physical system with spatial extent, we must also ask what happens at nonzero momentum—when the symmetry coordinate varies slowly from point to point.

Let $\theta(x)$ parameterize the position on the symmetry orbit at each spatial point. A uniform rotation $\theta = \text{const}$ costs no energy (Goldstone theorem). But a spatially varying $\theta(x)$ does cost energy, because the fields at neighboring points are no longer aligned.

The most general effective action for θ consistent with the symmetry, to leading order in derivatives, is

$$S_{\text{hydro}}[\theta] = \frac{1}{2} \int \mathrm{d}\tau \, \mathrm{d}^d x \left[\chi \, (\partial_\tau \theta)^2 + \rho_s \, (\nabla \theta)^2 \right] + \cdots , \tag{25.1}$$

where χ is a **temporal susceptibility** (or compressibility) and ρ_s is the **stiffness** (or superfluid density, or spin stiffness, depending on context). Both can be computed from the microscopic theory.

The Hessian in Fourier space is

$$K_\theta(\omega_n, \mathbf{q}) = \chi \omega_n^2 + \rho_s \mathbf{q}^2.$$

At $\omega_n = 0$ and $\mathbf{q} = 0$, this vanishes: the Goldstone theorem. At nonzero momentum, the phase propagator is

$$D_\theta(\omega_n, \mathbf{q}) = \frac{1}{\chi \omega_n^2 + \rho_s \mathbf{q}^2}.$$

After analytic continuation ($\mathrm{i}\omega_n \to \omega + \mathrm{i}0^+$), the pole condition gives

$$\omega^2 = \frac{\rho_s}{\chi} \, q^2 = c^2 q^2, \tag{25.2}$$

where $c = \sqrt{\rho_s / \chi}$ is the speed of the collective mode.

Chapter 26

Soft Modes from Conservation Laws

Not all soft modes are Goldstone modes. A conserved density $\partial_t n + \nabla \cdot \mathbf{j} = 0$ constrains the small-q, small-ω inverse covariance. Fick's law $\mathbf{j} = -D_{\text{diff}} \nabla n$ gives **diffusion**: $\omega = -\mathrm{i} D_{\text{diff}} q^2$. Adding a restoring pressure force gives **sound**: $\omega = \pm cq - \mathrm{i}\Gamma q^2$. Both are gapless because a uniform density perturbation cannot relax—total particle number is conserved.

Mathematical structure	Physics terminology
Small precision eigenvalue	Soft mode
Null direction from spontaneous symmetry breaking	Goldstone mode
Slow direction from a conservation law	Hydrodynamic mode
Pole of covariance at real frequency	Propagating collective excitation
Pole at imaginary frequency	Relaxation or diffusion

26.1 Deriving Hydrodynamic Modes from Conservation Laws

Let us derive the diffusion and sound poles from the structure of the inverse propagator, in the spirit of this book's framework.

Diffusion from a conserved density

Consider a conserved density $n(x, t)$ satisfying the continuity equation $\partial_t n + \nabla \cdot \mathbf{j} = 0$. The continuity equation is an exact identity that constrains the structure of the density-density response function $\chi(q, \omega) = \langle n(q, \omega)\, n(-q, -\omega) \rangle_c$.

In Fourier space, the continuity equation relates the density response to the current response:

$$-\mathrm{i}\omega\, \chi_{nn}(q, \omega) + \mathrm{i}q_i\, \chi_{j_i n}(q, \omega) = 0.$$

This constrains $\chi_{nn}(0, \omega)$: at $q = 0$, the conservation law forces $\omega\, \chi_{nn}(0, \omega) = 0$, so either $\omega = 0$ or χ_{nn} vanishes. This means the inverse susceptibility $K_{nn}(q = 0, \omega = 0) = \chi_{nn}^{-1}(0, 0)$ must be finite (the susceptibility itself is finite), and the mode frequency must vanish at $q = 0$. Conservation laws guarantee gapless modes.

Now add the constitutive relation (Fick's law): $\mathbf{j} = -D_{\text{diff}} \nabla n$ for a system without a restoring force. Substituting into the continuity equation:

$$\partial_t n = D_{\text{diff}} \nabla^2 n,$$

which in Fourier space gives $(-\mathrm{i}\omega + D_{\text{diff}}q^2)n(q,\omega) = 0$. The inverse propagator for the density mode is therefore

$$K_{\text{diff}}(\omega, q) = -\mathrm{i}\omega\chi + D_{\text{diff}}\chi q^2, \tag{26.1}$$

where $\chi = \partial n/\partial\mu$ is the static compressibility (a thermodynamic coefficient). The mode condition $K_{\text{diff}} = 0$ gives $\omega = -\mathrm{i}D_{\text{diff}}q^2$: a purely imaginary frequency, corresponding to exponential relaxation with rate $D_{\text{diff}}q^2$. At $q = 0$, the relaxation rate vanishes: the uniform density perturbation does not decay (it is exactly conserved).

Sound from a conserved density with a restoring force

If the density perturbation creates a pressure response $\delta P = (\partial P/\partial n)\delta n \equiv c_0^2\delta n$, the equation of motion acquires a restoring force. The linearized Euler equation for the momentum density gives

$$\partial_t j_i = -\nabla_i P + \eta\nabla^2 j_i + (\zeta + \eta/3)\nabla_i(\nabla\cdot\mathbf{j}),$$

where η and ζ are the shear and bulk viscosities. Together with the continuity equation, this gives (for longitudinal sound, projecting onto the $\hat{q}$ direction):

$$\partial_t^2 n = c_0^2\nabla^2 n + \Gamma_s\nabla^2\partial_t n,$$

where $\Gamma_s = (\zeta + 4\eta/3)/(mn_0)$ is the sound attenuation coefficient. In Fourier space, the inverse propagator is

$$K_{\text{sound}}(\omega, q) = \chi\omega^2 - \rho_s q^2 + \mathrm{i}\chi\Gamma_s\omega q^2, \tag{26.2}$$

where $\rho_s = \chi c_0^2$ is the stiffness. The mode condition $K_{\text{sound}} = 0$ gives the dispersion relation

$$\omega^2 + \mathrm{i}\Gamma_s q^2\omega - c_0^2 q^2 = 0,$$

with solution $\omega = \pm c_0 q - \frac{\mathrm{i}}{2}\Gamma_s q^2 + O(q^3)$. This is a propagating mode with speed c_0 and damping rate $\Gamma = \frac{1}{2}\Gamma_s q^2$.

Key point: the sound mode exists because the conservation of both particle number and momentum forces the inverse propagator to have the structure $\chi\omega^2 - \rho_s q^2 + \cdots$, which vanishes at $\omega = q = 0$. No microscopic weak-coupling assumption is needed. Even in a strongly interacting fluid where no quasiparticles exist, sound persists—its existence is guaranteed by symmetry (conservation laws), not by weak scattering.

Compare this with the Goldstone mode (Chapter 21), where the Hessian has a zero eigenvalue because of spontaneous symmetry breaking. Both mechanisms produce gapless modes, but for different reasons: Goldstone modes require a broken continuous symmetry; hydrodynamic modes require a conservation law. A system can have both (as in a superfluid, which has both a conserved particle number and a broken $U(1)$ symmetry, producing the Anderson–Bogoliubov sound mode).

26.2 Physical Background: Why Hydrodynamics Is Universal

Hydrodynamics is the most universal effective theory in all of physics. Its universality rests on a simple principle: conservation laws force certain modes to be gapless, regardless of the strength of interactions or the validity of any quasiparticle description.

Consider a system with a conserved charge $Q = \int n(x)\,\mathrm{d}^d x$, where n is the charge density. Conservation means $\partial_t n + \nabla\cdot\mathbf{j} = 0$ exactly, as an operator equation. This identity constrains the structure of the retarded density-density response function $\chi_R(\omega, q) = \langle n\, n\rangle_R$ at small ω and q.

The argument is as follows. The Ward identity for charge conservation (Chapter 34) gives

$$\omega\,\chi_{nn}(\omega, q) = q_i\,\chi_{j_i n}(\omega, q).$$

At $q = 0$: $\omega\,\chi_{nn}(\omega, 0) = 0$. For $\omega \neq 0$, this forces $\chi_{nn}(\omega, 0) = 0$—the uniform density does not fluctuate dynamically, because total charge is conserved. The inverse susceptibility $K(\omega, q)$ must therefore satisfy $K(0, 0) = 1/\chi_{\text{static}}$ (finite), and the mode frequency must vanish as $q \to 0$: the collective mode is **gapless**.

This is fundamentally different from a Goldstone mode, which is gapless because of spontaneous symmetry breaking. A hydrodynamic mode is gapless because of an exact conservation law. The conservation law is more robust: it holds at all temperatures, in all phases, and regardless of whether any symmetry is spontaneously broken.

The hierarchy of hydrodynamic modes

In a generic fluid (no broken symmetries, no long-range forces), the hydrodynamic modes are determined by the conservation laws alone:

Conserved particle number: gives a *diffusion* mode, $\omega = -\mathrm{i}Dq^2$. The diffusion constant D is a transport coefficient related to the conductivity by the Einstein relation $D = \sigma/\chi_n$.

Conserved momentum: combined with the equation of state $P(\rho)$, gives two *sound* modes $\omega = \pm c_s q - \mathrm{i}\Gamma q^2$. The sound speed $c_s = \sqrt{\partial P/\partial\rho}$ is a thermodynamic quantity; the attenuation Γ involves viscosities.

Conserved energy: gives a *thermal diffusion* mode, $\omega = -\mathrm{i}\kappa q^2/(c_p\rho)$, where κ is the thermal conductivity and c_p is the specific heat at constant pressure.

In total, a d-dimensional fluid with conserved particle number, momentum (d components), and energy has $d+2$ hydrodynamic modes: 2 propagating sound modes, $d-1$ transverse diffusion modes (shear), and 2 diffusive modes (thermal and particle diffusion, which may be coupled).

Hydrodynamic modes from the gap equation

How do hydrodynamic modes appear within the framework of this book? Through the Hessian of the effective action.

Consider a system where the collective field σ includes the local density $n(x)$. The effective action $S_{\text{eff}}[\sigma]$ inherits the conservation law: it is invariant under the global shift $n(x) \to n(x) + \epsilon$ (corresponding to total particle number conservation). At the saddle $\bar\sigma$, this invariance forces the Hessian $K_{nn}(q = 0, \omega = 0)$ to be finite but the mode frequency to vanish.

More concretely, the gap equation $\delta S_{\text{eff}}/\delta\sigma = 0$ determines the equilibrium density $\bar n$. The Hessian $K = \delta^2 S_{\text{eff}}/\delta\sigma^2$ at the saddle has the structure

$$K_{nn}(\omega, q) = \frac{1}{g_n} - \chi_0(\omega, q),$$

where $1/g_n$ is the bare inverse propagator for density fluctuations and χ_0 is the microscopic susceptibility. The gap equation $1/g_n = \chi_0(0, 0)$ ensures $K_{nn}(0, 0) = 0$—wait, this would imply a divergent static susceptibility, which is the criterion for a phase transition, not for a hydrodynamic mode.

The resolution is important: a hydrodynamic mode arises when the gap equation is *automatically* satisfied by a conservation law, rather than being a fine-tuned condition. The conservation law forces $K_{nn}(0, 0)$ to be zero (or, more precisely, forces the dynamical part of K to vanish at $\omega = q = 0$), without requiring any special value of the coupling g_n. This is the sense in which hydrodynamic modes are "protected" by conservation laws: they are not the result of a gap equation that might be disrupted by corrections, but a structural constraint on the form of the effective action.

In the person/society analogy: a hydrodynamic mode is like the conservation of total population. No social policy (no value of the "coupling constant") can make the total population change instantaneously. A density perturbation must spread by diffusion (migration) or propagate as a wave (a migration wave driven by economic forces). The "gap" for population dynamics is always zero, regardless of the strength of social interactions, because population is conserved.

The person/society analogy for hydrodynamic modes

In the person/society analogy, hydrodynamic modes correspond to quantities that are conserved in the society: total population, total wealth, total energy. A perturbation in the local population density (say, a sudden influx of people into one city) cannot disappear instantly—people do not vanish. The perturbation must spread out by diffusion (people moving to neighboring cities) or propagate as a wave (a migration wave). The diffusion constant is set by how mobile people are (the transport coefficient), not by the strength of social interactions. Even in a society with very strong social interactions (where no "quasiparticle" individual modes exist), diffusion of conserved quantities persists.

Part VII

Superconductivity as the Central Case Study

Superconductivity brings together every major theme of this book in a single, physically important application. The mathematical structure is rich enough to illustrate all the main ideas, but simple enough (at the mean-field level) to be worked out explicitly.

Chapter 27

Pairing as Latent-Variable Augmentation

Starting from an attractive interaction $-g\,\bar{\psi}_\uparrow\bar{\psi}_\downarrow\psi_\downarrow\psi_\uparrow$, introduce a complex pairing field $\Delta(x)$ coupled to $\psi_\downarrow\psi_\uparrow$ via HST. In Nambu notation $\Psi = (\psi_\uparrow, \bar{\psi}_\downarrow)^\top$, the action is quadratic in Ψ, with inverse Nambu propagator

$$\mathcal{G}^{-1}[\Delta] = \begin{pmatrix} G_0^{-1} & -\Delta \\ -\Delta^* & -\tilde{G}_0^{-1} \end{pmatrix},$$

where $G_0^{-1} = -\partial_\tau - \xi(-\mathrm{i}\nabla)$. For any fixed Δ, the fermion action is Gaussian in Ψ. Integrating out fermions exactly:

$$\boxed{S_{\text{eff}}[\Delta] = \frac{1}{g}\int |\Delta|^2 - \operatorname{Tr}\log \mathcal{G}^{-1}[\Delta] + \text{const.}} \tag{27.1}$$

This $\operatorname{Tr}\log$ is exact: fermions entered quadratically, and the Gaussian integral was performed without approximation.

Chapter 28

The BCS Gap Equation

28.1 The Saddle-Point Condition

At a uniform real saddle Δ_0, the stationarity condition requires $\frac{\partial S_{\text{eff}}}{\partial \Delta_0} = 0$. Using Jacobi's formula, this becomes

$$\frac{2\Delta_0}{g} = -\operatorname{Tr}(\mathcal{G}_{\Delta_0}\tau_1),$$

where $\tau_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ in Nambu space.

28.2 Evaluating the Nambu Propagator

The inverse Nambu propagator at the uniform saddle is

$$\mathcal{G}_{\Delta_0}^{-1}(k) = \begin{pmatrix} \mathrm{i}\omega_n - \xi_{\mathbf{k}} & -\Delta_0 \\ -\Delta_0 & \mathrm{i}\omega_n + \xi_{\mathbf{k}} \end{pmatrix},$$

where $\omega_n = (2n+1)\pi T$ are fermionic Matsubara frequencies and $\xi_{\mathbf{k}} = \mathbf{k}^2/(2m) - \mu$. Inverting the 2×2 matrix (using $M^{-1} = \frac{1}{ad-bc}\left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$):

$$\det \mathcal{G}_{\Delta_0}^{-1}(k) = -(\omega_n^2 + \xi_{\mathbf{k}}^2 + \Delta_0^2) = -(\omega_n^2 + E_{\mathbf{k}}^2),$$

where $E_{\mathbf{k}} = \sqrt{\xi_{\mathbf{k}}^2 + \Delta_0^2}$ is the **quasiparticle energy**. The Nambu trace picks out $\operatorname{tr}[\tau_1 \mathcal{G}_{\Delta_0}(k)] = -2\Delta_0/(\omega_n^2 + E_{\mathbf{k}}^2)$.

28.3 The Gap Equation

Substituting and using the standard Matsubara sum $T\sum_{\omega_n} 1/(\omega_n^2 + E^2) = \frac{1}{2E}\tanh(E/(2T))$, which at zero temperature gives $1/(2E)$:

$$\boxed{\frac{1}{g} = \int_{\mathbf{k}} \frac{1}{2E_{\mathbf{k}}} = \int_{\mathbf{k}} \frac{1}{2\sqrt{\xi_{\mathbf{k}}^2 + \Delta_0^2}}.} \tag{28.1}$$

This is the **BCS gap equation at zero temperature**. The solution has the non-perturbative form $\Delta_0 \sim \omega_D \exp(-1/(N_0 g))$, explaining why the gap cannot be seen in any finite order of perturbation theory.

Derivation of the Matsubara frequency sum

The sum $T\sum_{\omega_n} \frac{1}{\omega_n^2+E^2}$, where $\omega_n = (2n+1)\pi T$ are fermionic Matsubara frequencies, is a standard calculation in thermal field theory. We derive it here because it illustrates the contour techniques of Part XII and appears repeatedly.

Define $f(z) = \frac{1}{z^2+E^2} \cdot \frac{1}{e^{z/T}+1}$. The fermionic Matsubara sum is related to a contour integral:

$$T\sum_{n=-\infty}^{\infty} \frac{1}{\omega_n^2+E^2} = -\oint_{\mathcal{C}} \frac{dz}{2\pi \mathrm{i}} \frac{n_F(z)}{z^2+E^2},$$

where the contour $\mathcal{C}$ encircles all the fermionic Matsubara frequencies $z = \mathrm{i}\omega_n$ (which are the poles of $n_F(z) = 1/(e^{z/T}+1)$). The residue of $n_F(z)$ at $z = \mathrm{i}\omega_n$ is $-T$.

Deforming the contour to encircle instead the poles of $1/(z^2+E^2)$ at $z = \pm E$ (picking up a minus sign from the contour reversal):

$$T\sum_{n} \frac{1}{\omega_n^2+E^2} = n_F(E)\cdot\frac{1}{2E} + n_F(-E)\cdot\frac{-1}{2E}.$$

Using $n_F(-E) = 1 - n_F(E)$:

$$\boxed{T\sum_{\omega_n} \frac{1}{\omega_n^2+E^2} = \frac{1}{2E}\left[1-2n_F(E)\right] = \frac{1}{2E}\tanh\frac{E}{2T}.} \tag{28.2}$$

At $T=0$: $n_F(E)\to 0$ (for $E>0$), so the sum gives $1/(2E)$, recovering the zero-temperature result. At high T: $\tanh(E/(2T)) \approx E/(2T)$, so the sum gives $1/(4T)$—independent of E, reflecting the classical (high-temperature) limit.

The finite-temperature BCS gap equation

At temperature $T>0$, the Matsubara sum in the gap equation gives

$$\boxed{\frac{1}{g} = \int_{\mathbf{k}} \frac{\tanh(E_{\mathbf{k}}/(2T))}{2E_{\mathbf{k}}}.} \tag{28.3}$$

As T increases from zero, $\tanh(E/(2T))$ decreases from 1, reducing the right-hand side. At a critical temperature T_c, the right-hand side becomes too small to sustain a nonzero Δ_0, and the gap vanishes continuously: $\Delta_0(T)\to 0$ as $T\to T_c^-$. At $T=T_c$, setting $\Delta_0=0$ (so $E_{\mathbf{k}} = |\xi_{\mathbf{k}}|$):

$$\frac{1}{g} = \int_{\mathbf{k}} \frac{\tanh(|\xi_{\mathbf{k}}|/(2T_c))}{2|\xi_{\mathbf{k}}|}.$$

Subtracting the zero-temperature gap equation ($1/g = \int_{\mathbf{k}} 1/(2E_{\mathbf{k}})$ with $E_{\mathbf{k}} = \sqrt{\xi_{\mathbf{k}}^2+\Delta_0^2}$) from this gives a relation between T_c and Δ_0. For weak coupling (where $\Delta_0 \ll \omega_D$, the Debye cutoff), the BCS theory gives the universal ratio

$$\frac{2\Delta_0}{k_B T_c} = \frac{\pi}{e^{\gamma_E}} \approx 3.528, \tag{28.4}$$

where γ_E is the Euler–Mascheroni constant. This is a quantitative prediction: the ratio of the zero-temperature gap to the critical temperature is a universal number, independent of the coupling constant g, the density of states, or the cutoff.

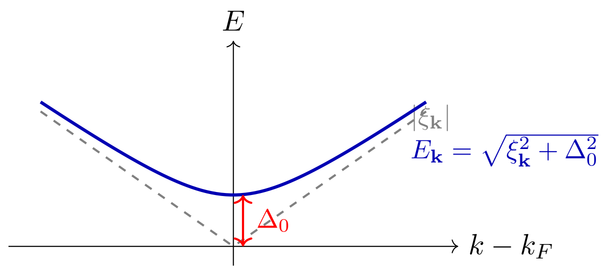
Near T_c, the gap vanishes as $\Delta_0(T) \propto \sqrt{1 - T/T_c}$—a mean-field square-root behavior characteristic of second-order phase transitions. This is the prediction of the saddle-point (mean-field) approximation. Fluctuation corrections (from the Gaussian integration over collective-field fluctuations in Part V) modify this behavior in low dimensions but do not change the qualitative picture in three dimensions.

Chapter 29

Gapped Quasiparticles and Soft Phase Fluctuations

29.1 The Quasiparticle Gap

The minimum quasiparticle energy is $E_{\min} = \Delta_0 > 0$, occurring at the Fermi surface where $\xi_{\mathbf{k}} = 0$. This is the **quasiparticle gap**: the minimum energy required to create a fermionic excitation.



The dashed line shows the normal-state dispersion $|\xi_{\mathbf{k}}|$, which vanishes at the Fermi surface. The solid curve shows the BCS dispersion $E_{\mathbf{k}} = \sqrt{\xi_{\mathbf{k}}^2 + \Delta_0^2}$, which has a minimum value $\Delta_0 > 0$: the gap.

29.2 Amplitude and Phase Fluctuations

Writing $\Delta(x) = [\Delta_0 + \rho(x)]e^{\mathrm{i}\theta(x)}$, the $U(1)$ symmetry of S_{eff} (invariance under $\Delta \to e^{\mathrm{i}\alpha}\Delta$) is broken by the saddle Δ_0. By the Goldstone theorem:

The **amplitude mode** ρ has $K_\rho(q = 0) > 0$—it is gapped, with mass of order $2\Delta_0$ (the "Higgs mode").

The **phase mode** θ has $K_\theta(0, 0) = 0$—it is gapless. At nonzero momentum: $K_\theta(\omega_n, \mathbf{q}) = \chi\omega_n^2 + \rho_s\mathbf{q}^2$, giving the Anderson–Bogoliubov sound mode $\omega = c_s q$.

> The pairing saddle simultaneously opens a gap Δ_0 in the fermionic quasiparticle covariance and produces a zero-Hessian direction in the collective phase covariance. These are not contradictory—they refer to different variables and different covariance operators.

Chapter 30

Gauge Coupling and the Fate of the Phase Mode

When the electrons are charged and couple to the electromagnetic gauge field A_μ, the effective action for (θ, A) expanded to quadratic order has a block Hessian

$$K = \begin{pmatrix} K_{\theta\theta} & K_{\theta A} \\ K_{A\theta} & K_{AA} \end{pmatrix}.$$

The effective precision for θ, after accounting for the gauge field via the Schur complement, is

$$K_\theta^{\text{eff}} = K_{\theta\theta} - K_{\theta A} K_{AA}^{-1} K_{A\theta}.$$

The off-diagonal coupling $K_{\theta A} \propto \rho_s q$ (from the gauge-covariant gradient $\nabla\theta - eA$) and $K_{AA}^{-1} \propto e^2/q^2$ (the Coulomb propagator) combine to give a constant mass term:

$$K_{\theta A} K_{AA}^{-1} K_{A\theta} \sim \rho_s^2 e^2 = \text{const}.$$

This adds a gap to the phase mode: the plasma frequency $\omega_p^2 \propto ne^2/m$. The former Goldstone mode acquires a mass. This is the **Anderson–Higgs mechanism**.

The phrase "the gauge field eats the Goldstone boson" is a memorable story about the spectral analysis of a coupled block precision operator. The story is vivid and useful, but it should not be confused with the derivation, which is the Schur complement calculation above.

Let us see this Schur complement calculation in more detail. Write the quadratic action for the coupled $(\theta, \mathbf{A})$ fluctuations:

$$S^{(2)} = \frac{1}{2} \int \frac{d^3q}{(2\pi)^3} T \sum_{\omega_n} \begin{pmatrix} \theta(-q) & \mathbf{A}(-q) \end{pmatrix} \begin{pmatrix} \chi\omega_n^2 + \rho_s q^2 & -2e\rho_s \mathrm{i}q \\ -2e\rho_s \mathrm{i}q & q^2 + 4e^2\rho_s \end{pmatrix} \begin{pmatrix} \theta(q) \\ \mathbf{A}(q) \end{pmatrix},$$

where we have shown only the longitudinal component of $\mathbf{A}$ for simplicity. The off-diagonal coupling $K_{\theta A} = -2e\rho_s \mathrm{i}q$ arises from the gauge-covariant derivative $|\nabla\theta - 2eA|^2$ in the phase action.

The effective phase precision, after integrating out $\mathbf{A}$ via the Schur complement, is

$$\begin{aligned} K_\theta^{\text{eff}}(\omega_n, q) &= K_{\theta\theta} - K_{\theta A} K_{AA}^{-1} K_{A\theta} \\ &= (\chi\omega_n^2 + \rho_s q^2) - \frac{(2e\rho_s q)^2}{q^2 + 4e^2\rho_s} \\ &= \chi\omega_n^2 + \rho_s q^2 \cdot \frac{q^2}{q^2 + 4e^2\rho_s}. \end{aligned}$$

At $q = 0$: $K_\theta^{\text{eff}}(\omega_n, 0) = \chi\omega_n^2$, which does not vanish at $\omega_n = 0$—but this is because the static $q = 0$ mode has been absorbed into a gauge transformation (it is unphysical). At small but nonzero q: $K_\theta^{\text{eff}} \approx \chi\omega_n^2 + \rho_s q^4/(4e^2\rho_s) = \chi\omega_n^2 + q^4/(4e^2)$, which has a q^4 dispersion—not the q^2 of the original Goldstone mode.

Meanwhile, the effective gauge-field precision acquires a mass from integrating out θ:

$$K_A^{\text{eff}} = (q^2 + 4e^2\rho_s) - \frac{(2e\rho_s q)^2}{\chi\omega_n^2 + \rho_s q^2}.$$

At $\omega_n = q = 0$: $K_A^{\text{eff}}(0,0) = 4e^2\rho_s \equiv \omega_p^2/c^2$. The gauge field has acquired a mass $m_A^2 = 4e^2\rho_s$: this is the **Meissner mass**. The penetration depth is $\lambda_L = 1/\sqrt{4e^2\rho_s}$, and magnetic fields are exponentially screened inside the superconductor.

This is the Anderson–Higgs mechanism as a Schur complement: the off-diagonal coupling between the Goldstone mode (θ) and the gauge field (A) transfers the zero-Hessian eigenvalue from θ to A, converting the massless Goldstone mode into a massive gauge field. The “eating” of the Goldstone boson is mathematically the off-diagonal block $K_{\theta A}$ connecting the two sectors.

Part VIII

Additional Worked Examples

The framework developed in Parts I–VII applies to a wide range of physical systems. This part works through four additional examples, chosen to illustrate different facets of the general structure. Each example follows the same logical sequence: define the model, introduce a latent field, integrate out the microscopic variables, derive the effective action, find the saddle, compute the Hessian, identify the collective modes, and tell the physical story.

Chapter 31

Scalar ϕ^4 Theory and Auxiliary Mass Fields

31.1 The Model and Direct Perturbation Theory

The scalar ϕ^4 theory has action

$$S[\phi] = \int \left[\frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m_0^2\phi^2 + \frac{\lambda}{4!}\phi^4 \right] \mathrm{d}^d x.$$

In Fourier space, the bare propagator is $C_0(k) = 1/(k^2 + m_0^2)$, and the quartic term is the non-Gaussian interaction.

31.2 The Tadpole from Perturbation Theory

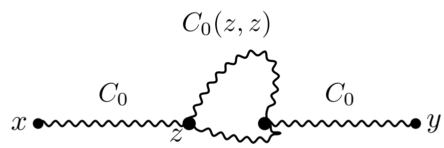
The first perturbative correction to the two-point function $\langle\phi(x)\phi(y)\rangle$ at order λ comes from

$$-\frac{\lambda}{4!}\int \mathrm{d}^d z \; \langle\phi(x)\phi(y)\phi(z)^4\rangle_0,$$

where $\langle\cdot\rangle_0$ denotes the Gaussian ($\lambda = 0$) expectation. By Wick's theorem, the six-field correlator produces several pairings. The connected contribution has two of the four $\phi(z)$ fields paired with each other (forming a closed loop at z) and the remaining two paired with $\phi(x)$ and $\phi(y)$. The result is proportional to

$$\lambda\, C_0(x,z)\, C_0(z,y)\, C_0(z,z).$$

The factor $C_0(z,z) = \int_k C_0(k)$ is the equal-point propagator—a loop. Diagrammatically, this is a single propagator line from x to z, a loop at z, and another propagator from z to y: a "tadpole" hanging off the propagator.



In momentum space, the self-energy correction is

$$\Sigma_{\text{tadpole}} = \frac{\lambda}{2}\int_k \frac{1}{k^2 + m_0^2}.$$

This is momentum-independent—it is a pure mass correction. The effective mass becomes $m_{\text{eff}}^2 = m_0^2 + \Sigma_{\text{tadpole}}$.

We pause to emphasize: the leading self-energy in ϕ^4 theory is a tadpole, not a two-propagator bubble. The resolution is that "bubble" refers to a specific algebraic structure (a trace of two propagators), while "self-energy" refers to the full precision correction. Which diagrams contribute depends on the theory.

31.3 The HST Approach

Introducing an auxiliary field σ coupled to ϕ^2:

$$S_{\text{eff}}[\sigma] = \frac{3}{\lambda}\int \sigma^2 + \frac{1}{2}\operatorname{Tr}\log(-\nabla^2 + m_0^2 + \sigma).$$

The saddle equation for a uniform $\bar{\sigma}$ with $m^2 = m_0^2+\bar{\sigma}$ is a self-consistent mass equation that resums the tadpole. The collective Hessian $K(q) = 6/\lambda + \Pi(q)$ exhibits **critical softening**: as $m^2 \to 0$, $\Pi(0)$ diverges in low dimensions ($d \leq 4$), $K(0) \to 0$, and the collective covariance $D = K^{-1}$ diverges—signaling a phase transition.

Chapter 32

The $O(N)$ Model and Large-N

32.1 The Model

The $O(N)$ model generalizes ϕ^4 to a vector field $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ with $O(N)$ symmetry:

$$S[\boldsymbol{\phi}] = \int \left[\frac{1}{2}(\nabla\boldsymbol{\phi})^2 + \frac{1}{2}m_0^2|\boldsymbol{\phi}|^2 + \frac{\lambda}{4N}(|\boldsymbol{\phi}|^2)^2\right] \mathrm{d}^d x.$$

The coupling is scaled as $\lambda/(4N)$ so that the free energy is extensive in N.

32.2 Auxiliary Field and Large-N Concentration

Introducing a scalar auxiliary field σ coupled to $|\boldsymbol{\phi}|^2$ via HST, each of the N components ϕ_a is an independent Gaussian with precision $-\nabla^2 + m_0^2 + \sigma$. The effective action is

$$S_{\text{eff}}[\sigma] = \frac{N}{4\lambda}\int \sigma^2 + \frac{N}{2}\operatorname{Tr}\log(-\nabla^2 + m_0^2 + \sigma).$$

Every term is proportional to N: $S_{\text{eff}} = N \cdot s_{\text{eff}}[\sigma]$. As $N \to \infty$, the integral $\int \mathcal{D}\sigma\, e^{-N s_{\text{eff}}}$ is dominated by the saddle, with corrections suppressed by powers of $1/N$. This is a beautiful example of the general mechanism discussed in Chapter 10: the prefactor N controls the sharpness of the saddle.

32.3 The Large-N Saddle and Mass Gap

The saddle equation for a uniform $\bar{\sigma}$ is

$$\frac{\bar{\sigma}}{2\lambda} = -\frac{1}{2}\int_k \frac{1}{k^2 + m_0^2 + \bar{\sigma}}.$$

Defining the physical mass $m^2 = m_0^2 + \bar{\sigma}$:

$$\frac{m^2 - m_0^2}{2\lambda} = -\frac{1}{2}\int_k \frac{1}{k^2 + m^2}.$$

This is a self-consistent mass equation, identical in structure to the ϕ^4 gap equation up to numerical factors. It determines the mass gap m—the inverse correlation length—as a function of the bare parameters. When $m_0^2 > 0$, the solution has $m^2 > 0$: the system is in the disordered (symmetric) phase with all N field components having equal mass. When m_0^2 is sufficiently negative, the symmetric saddle becomes unstable, signaling the onset of a broken-symmetry phase.

The Mermin–Wagner constraint. An important caveat: in dimensions $d \leq 2$, continuous symmetry breaking is forbidden for systems with short-range interactions (the Mermin–Wagner–Hohenberg–Coleman theorem). The saddle-point analysis, which predicts symmetry breaking for sufficiently negative m_0^2 in any dimension, is an artifact of the mean-field approximation. The $1/N$ corrections (discussed below) restore the correct behavior: in $d = 2$, infrared fluctuations of the Goldstone modes destroy long-range order, and the correlation length remains finite at any temperature. This does not mean the saddle-point analysis is useless in $d = 2$—it correctly captures the short-distance physics and the structure of the effective action—but its prediction of a phase transition must be interpreted with care. In $d = 3$, the saddle prediction of a broken phase is qualitatively correct.

32.4 $1/N$ Corrections

The leading correction beyond the saddle is the Gaussian fluctuation of σ—the Hessian and its inverse. The σ-Hessian is $K(q) = 1/(2\lambda) + \frac{1}{2}\Pi(q)$, where $\Pi(q) = \int_k G_m(k+q)G_m(k)$. Its effects on the microscopic propagators enter at order $1/N$. Higher vertices produce corrections at order $1/N^2$ and beyond.

32.5 The Broken Phase

In the broken phase, $\langle \boldsymbol{\phi} \rangle = (v, 0, \ldots, 0)$ for some $v > 0$, breaking $O(N) \to O(N-1)$. The fluctuations decompose into one gapped radial (longitudinal) mode and $N-1$ gapless angular (transverse) Goldstone modes. The number of Goldstone modes equals $\frac{N(N-1)}{2} - \frac{(N-1)(N-2)}{2} = N-1$.

Chapter 33

The Electron Gas, Screening, and Plasmons

33.1 HST in the Density Channel

For electrons interacting through $V(\mathbf{q}) = 4\pi e^2/\mathbf{q}^2$, introduce a scalar field φ coupled to the density n:

$$S_{\text{eff}}[\varphi] = \frac{1}{2}\int \varphi\, V^{-1}\varphi - 2\ \mathrm{Tr}\log(G_0^{-1} + \mathrm{i}\varphi) + \text{const.}$$

The factor of i in the coupling arises because the Coulomb interaction is repulsive.

33.2 Screening and Plasmons

The Hessian gives the screened interaction:

$$D(q) = \frac{V(q)}{1 + V(q)\Pi(q)} = \frac{V(q)}{\varepsilon(q)},$$

where $\varepsilon(q) = 1 + V(q)\Pi(q)$ is the **dielectric function**. At small $\mathbf{q}$: $\varepsilon(q) \approx 1 + q_{\text{TF}}^2/q^2$ with $q_{\text{TF}} = \sqrt{4\pi e^2 N_0}$ (Thomas–Fermi wavevector), giving a Yukawa potential $D(r) \propto e^{-q_{\text{TF}} r}/r$.

To extract the screening length, consider the static ($\omega = 0$) limit at small $\mathbf{q}$. The static polarization approaches a constant: $\Pi(0, \mathbf{q} \to 0) \to -N_0$ (the density of states at the Fermi energy, with a sign from conventions). Therefore

$$\varepsilon(\mathbf{q}) \approx 1 + \frac{4\pi e^2}{q^2} \cdot N_0 = 1 + \frac{q_{\text{TF}}^2}{q^2},$$

where $q_{\text{TF}} = \sqrt{4\pi e^2 N_0}$ is the **Thomas–Fermi screening wavevector**. The screened interaction becomes

$$D(\mathbf{q}) = \frac{4\pi e^2}{q^2 + q_{\text{TF}}^2},$$

which is a **Yukawa potential** in real space: $D(r) \propto e^{-q_{\text{TF}} r}/r$. The long-range Coulomb interaction has been screened to a finite range $r_{\text{TF}} = 1/q_{\text{TF}}$.

The Lindhard function: explicit evaluation

Let us compute the free-fermion polarization bubble $\Pi_0(q,\omega_n)$ explicitly, as it is the prototype for every bubble calculation in this book. We need to evaluate

$$\Pi_0(\mathrm{i}\omega_n,\mathbf{q}) = -2\,T\sum_{\nu_m}\int \frac{d^3k}{(2\pi)^3}\,G_0(\mathrm{i}\nu_m,\mathbf{k})\,G_0(\mathrm{i}\nu_m+\mathrm{i}\omega_n,\mathbf{k}+\mathbf{q}),$$

where $G_0(\mathrm{i}\nu_m,\mathbf{k}) = 1/(\mathrm{i}\nu_m - \xi_\mathbf{k})$ is the free fermion propagator, $\xi_\mathbf{k} = k^2/(2m_e) - \mu$ is the kinetic energy measured from the chemical potential, $\nu_m = (2m+1)\pi T$ are fermionic Matsubara frequencies, and the factor of 2 is for spin.

Step 1: Matsubara sum. We need $T\sum_{\nu_m}\frac{1}{(\mathrm{i}\nu_m-\xi_\mathbf{k})(\mathrm{i}\nu_m+\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}})}$. Use partial fractions:

$$\frac{1}{(\mathrm{i}\nu_m-\xi_\mathbf{k})(\mathrm{i}\nu_m+\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}})} = \frac{1}{\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}}+\xi_\mathbf{k}}\left[\frac{1}{\mathrm{i}\nu_m-\xi_\mathbf{k}} - \frac{1}{\mathrm{i}\nu_m+\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}}}\right].$$

Now use the standard Matsubara sum (derived in detail in Chapter 41):

$$T\sum_{\nu_m}\frac{1}{\mathrm{i}\nu_m-\xi} = n_F(\xi),$$

where $n_F(\xi) = 1/(e^{\xi/T}+1)$ is the Fermi–Dirac distribution. Also, $T\sum_{\nu_m}\frac{1}{\mathrm{i}\nu_m+\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}}} = n_F(\xi_{\mathbf{k}+\mathbf{q}})$ (since the bosonic frequency ω_n shifts the sum but does not change its value for the Fermi function). Therefore:

$$T\sum_{\nu_m} G_0(\mathrm{i}\nu_m,\mathbf{k})\,G_0(\mathrm{i}\nu_m+\mathrm{i}\omega_n,\mathbf{k}+\mathbf{q}) = \frac{n_F(\xi_\mathbf{k})-n_F(\xi_{\mathbf{k}+\mathbf{q}})}{\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}}+\xi_\mathbf{k}}.$$

Step 2: The Lindhard function. The polarization is

$$\boxed{\Pi_0(\mathrm{i}\omega_n,\mathbf{q}) = -2\int\frac{d^3k}{(2\pi)^3}\frac{n_F(\xi_\mathbf{k})-n_F(\xi_{\mathbf{k}+\mathbf{q}})}{\mathrm{i}\omega_n-\xi_{\mathbf{k}+\mathbf{q}}+\xi_\mathbf{k}}.} \tag{33.1}$$

This is the **Lindhard function**. At zero temperature, $n_F(\xi) = \theta(-\xi)$ (a step function at the Fermi surface), and the numerator $n_F(\xi_\mathbf{k}) - n_F(\xi_{\mathbf{k}+\mathbf{q}})$ is nonzero only when one momentum is inside the Fermi sea and the other is outside.

Step 3: Static limit. Setting $\omega_n = 0$:

$$\Pi_0(0,\mathbf{q}) = -2\int\frac{d^3k}{(2\pi)^3}\frac{n_F(\xi_\mathbf{k})-n_F(\xi_{\mathbf{k}+\mathbf{q}})}{\xi_\mathbf{k}-\xi_{\mathbf{k}+\mathbf{q}}}.$$

At $T=0$ and small q, expanding $\xi_{\mathbf{k}+\mathbf{q}} \approx \xi_\mathbf{k} + \mathbf{v}_\mathbf{k}\cdot\mathbf{q}$ (where $\mathbf{v}_\mathbf{k} = \nabla_\mathbf{k}\xi_\mathbf{k} = \mathbf{k}/m_e$ is the velocity), the numerator becomes $-\mathbf{v}_\mathbf{k}\cdot\mathbf{q}\,\delta(\xi_\mathbf{k})\cdot(-1)$, and the ratio gives $\delta(\xi_\mathbf{k})$. The integral over $\mathbf{k}$ then picks out the Fermi surface:

$$\Pi_0(0,\mathbf{q}\to 0) = -2\int\frac{d^3k}{(2\pi)^3}\,\delta(\xi_\mathbf{k}) = -N_0,$$

where $N_0 = m_e k_F/\pi^2$ is the single-spin density of states at the Fermi energy. This confirms the Thomas–Fermi result: $q_{\mathrm{TF}}^2 = 4\pi e^2 N_0$.

Order of limits. The static limit ($\omega\to 0$ first, then $q\to 0$) and the uniform limit ($q\to 0$ first, then $\omega\to 0$) give different results for the polarization. The *static* limit $\Pi(0,q\to 0) = -N_0$ gives the compressibility and screening. The *dynamic* (or uniform) limit $\Pi(\omega\to 0, q=0) = 0$ (which vanishes because at $q = 0$ no particle-hole excitations exist). This non-commutativity of limits is a general feature of response functions and has physical consequences: the static dielectric function $\varepsilon(0, q) = 1 + V(q)N_0$ gives screening, while the dynamic dielectric function $\varepsilon(\omega, 0) = 1 - \omega_p^2/\omega^2$ gives plasma oscillations. These probe different physics (charge rearrangement vs. collective oscillation), and the different limits reflect this.

For general q, the Lindhard function at $T = 0$ evaluates to

$$\Pi_0(0, q) = -N_0 \left[\frac{1}{2} + \frac{1 - x^2}{4x} \log \left| \frac{1 + x}{1 - x} \right| \right], \qquad x = \frac{q}{2k_F}. \tag{33.2}$$

This function equals $-N_0$ at $q = 0$, decreases monotonically for $q < 2k_F$, has a logarithmic singularity in its derivative at $q = 2k_F$ (the **Kohn anomaly**—a singularity arising because the Fermi surface has a sharp edge), and falls off as $\sim -N_0 k_F^2/(3q^2)$ for $q \gg 2k_F$.

Landau damping

The imaginary part of the retarded polarization, obtained by analytic continuation $\mathrm{i}\omega_n \to \omega + \mathrm{i}0^+$, is

$$\operatorname{Im} \Pi_0^R(\omega, \mathbf{q}) = -\pi \int \frac{d^3k}{(2\pi)^3} \left[n_F(\xi_{\mathbf{k}}) - n_F(\xi_{\mathbf{k+q}}) \right] \delta(\omega - \xi_{\mathbf{k+q}} + \xi_{\mathbf{k}}).$$

The delta function enforces energy conservation: a particle-hole pair with momentum transfer $\mathbf{q}$ absorbs energy $\omega = \xi_{\mathbf{k+q}} - \xi_{\mathbf{k}}$. For small ω and q, this is satisfied when $\omega \approx \mathbf{v}_F \cdot \mathbf{q}$, where $\mathbf{v}_F$ is the Fermi velocity. The integration over the direction of $\mathbf{k}$ on the Fermi surface gives

$$\operatorname{Im} \Pi_0^R(\omega, q) \approx -\pi N_0 \frac{\omega}{v_F q} \quad \text{for } |\omega| < v_F q.$$

This is **Landau damping**: a collective mode (such as a plasmon) with frequency ω and wavevector q can decay into particle-hole pairs whenever $|\omega| < v_F q$—the region in (ω, q)-space where individual particle-hole excitations exist. The imaginary part of Π gives the rate of this decay.

The plasmon dispersion $\omega_p(q)$ lies above the particle-hole continuum at small q (since $\omega_p \gg v_F q$ at small q), so the plasmon is undamped at long wavelengths. As q increases, $\omega_p(q)$ enters the continuum and acquires a damping rate $\Gamma(q) \propto \operatorname{Im} \Pi_0^R(\omega_p, q)$, which grows until the plasmon is completely overdamped. This is the physical origin of the Landau damping phenomenon.

Three notions of mass

Before discussing Debye screening in detail, we must distinguish three notions of "mass" that are routinely conflated in physics discussions:

Notion 1: the Lagrangian mass parameter m_0. This is the coefficient of ϕ^2 (or $A_\mu A^\mu$) in the action. It is a *natural parameter* of the exponential family $p[\phi] \propto e^{-S}$. For the photon in QED, $m_0 = 0$: the action has no mass term because such a term would break gauge invariance.

Notion 2: the pole mass m_{phys}. This is the location of the pole of the exact vacuum propagator: $G(k) \sim 1/(k^2 - m_{\text{phys}}^2)$ near the pole. It is a *mean parameter*—an observable extracted from the correlator, including all quantum corrections. For the photon in vacuum, $m_{\text{phys}} = 0$, protected by the Ward identity.

Notion 3: the screening mass m_D. In a medium (a plasma, a crystal, a condensate), the propagator is modified by the medium's response. The in-medium propagator has poles at $k^2 = m_D^2$, where m_D depends on the medium's properties (temperature, density, composition). This is the *Debye mass.*

The Debye mass is not a violation of gauge invariance. It does not arise from a mass term in the Lagrangian. It arises from the polarization of the medium: the collective rearrangement of charges screens the long-range interaction, and the screening length $\lambda_D = 1/m_D$ sets the range of the effective interaction.

Debye mass and its generalizations

The screening wavevector q_{TF} is one instance of a general concept that appears across many areas of physics under the name **Debye mass** m_D. The mathematical structure is always the same: a long-range interaction $V(q) \sim 1/q^2$ is screened by the static polarization $\Pi(0, \mathbf{q} \to 0)$ to produce a screened interaction $D(q) = V(q)/(1 + V(q)\Pi(0))$, which has a pole at $q^2 = -m_D^2$ with

$$m_D^2 = \lim_{\mathbf{q}\to 0} V(\mathbf{q})\,\Pi(0,\mathbf{q}). \tag{33.3}$$

In different contexts, m_D takes different values because the static polarization depends on the physics of the medium:

In a **degenerate electron gas** at zero temperature, $\Pi(0) \to N_0$ (the density of states at the Fermi surface), giving $m_D = q_{\mathrm{TF}} = \sqrt{4\pi e^2 N_0}$. This is the Thomas–Fermi result derived above.

In a **classical plasma** at temperature T with density n and charge e, $\Pi(0) \to ne^2/(k_B T)$ (from the Boltzmann distribution), giving the original **Debye screening mass** $m_D = \sqrt{4\pi n e^2/(k_B T)}$. The screening length $\lambda_D = 1/m_D$ is the **Debye length**—the scale beyond which charges in a plasma are effectively screened.

In **finite-temperature QED**, the photon acquires a thermal mass $m_D^2 \propto e^2 T^2$ from the polarization of the electron-positron plasma. This is an electric screening mass; it gaps the longitudinal (Coulomb) component of the photon propagator but not the transverse (magnetic) component—a subtlety related to the different infrared behavior of electric and magnetic screening.

In **QCD at high temperature**, the gluon acquires a Debye mass $m_D^2 \propto g^2 T^2$ from the polarization of the quark-gluon plasma. Electric (color-Coulomb) interactions are screened at the scale $1/m_D$, which plays a central role in the physics of quark-gluon plasma and in the suppression of heavy quark bound states (the Matsui–Satz mechanism of J/ψ suppression).

In every case, the Debye mass has the same mathematical origin: it is the infrared limit of the polarization bubble, inserted into the dressed collective propagator $D = (V^{-1} + \Pi)^{-1}$. It converts a massless (long-range) bare interaction into a massive (short-range) screened interaction. The formula $m_D^2 = V(0) \cdot \Pi(0)$ is a direct application of the Hessian structure $K = D_0^{-1} + \Pi$ from Chapter 15, evaluated at zero momentum.

The Ward identity: what it protects and what it doesn't

A natural question arises: if the photon is massless ($m_\gamma = 0$) because gauge invariance forbids a mass term, how can it acquire a Debye mass in a plasma? The resolution involves a careful distinction.

The Ward identity $k^\mu \Pi_{\mu\nu} = 0$ holds at *all* temperatures. In vacuum ($T = 0$), Lorentz invariance constrains the polarization tensor to the form $\Pi_{\mu\nu} \propto (k^2 g_{\mu\nu} - k_\mu k_\nu)$, which forces $\Pi_{\mu\nu}(0) = 0$: no vacuum photon mass.

In a medium ($T > 0$), Lorentz invariance is broken by the medium's rest frame. The polarization tensor decomposes into longitudinal and transverse components:

$$\Pi_{\mu\nu}(k) = \Pi_L(k)\, P^L_{\mu\nu} + \Pi_T(k)\, P^T_{\mu\nu},$$

where P^L and P^T are projectors with respect to $\mathbf{k}$ and the medium rest frame. The Ward identity still holds, but it no longer forces both components to vanish at $k = 0$:

$$\Pi_L(k_0 = 0, \mathbf{k} \to 0) = m_D^2 \neq 0, \qquad \Pi_T(k_0 = 0, \mathbf{k} \to 0) = 0.$$

Only the *longitudinal* (electric) mode acquires a static screening mass. The transverse (magnetic) modes are not screened at leading order—this is the *Linde problem* of hot QCD, and it reflects a deep asymmetry between electric and magnetic screening.

Three levels of description

The Debye mass fits into a three-level hierarchy that organizes the different "masses" appearing in a field theory:

Level	**What it is**	**How computed**	**Mass?**
1. Microscopic	Matter field mass m	Read from the action	Lagrangian parameter
2. Screening	Induced propagator of σ	Connected $\langle\sigma\sigma\rangle_c$	$m_D > 0$ (Debye)
3. Saddle-point	Classical profile $\bar{\sigma}$	EM / saddle at $N \to \infty$	No propagation

Level 3 is classical physics (Coulomb's law, Einstein's equation, sound). It exists at $N = \infty$ and does not require quantum mechanics. Level 2 is quantum: the screening mass is visible only in the connected correlator, requires finite N, and has amplitude $O(1/N)$. The Debye mass lives at Level 2.

At finite frequency, the collective mode condition $\varepsilon(\omega, \mathbf{q}) = 0$ gives the **plasma frequency** $\omega_p^2 = 4\pi n e^2/m$ at $q = 0$. The plasmon is a *gapped* collective mode: its frequency does not vanish as $q \to 0$, because the Coulomb restoring force provides finite energy even at zero wavevector. At nonzero q, the plasmon dispersion acquires a positive correction: $\omega^2(q) = \omega_p^2 + \alpha q^2 + \cdots$. At large enough q, the plasmon enters the particle-hole continuum (where $\operatorname{Im}\Pi \neq 0$) and decays into individual electron-hole pairs—a process called **Landau damping**.

Chapter 34

Magnetism and Spin Waves

34.1 HST in the Spin Channel

For a Hubbard interaction $-U \int n_\uparrow n_\downarrow$, introduce a vector field $\mathbf{h}$ coupled to the magnetization density. The **Stoner saddle** $\bar{\mathbf{h}} = h_0 \hat{z}$ produces exchange splitting: up-spin and down-spin bands are split by $2h_0$. The saddle condition is the Stoner criterion $UN_0 > 1$.

34.2 Longitudinal and Transverse Modes

The saddle $\bar{\mathbf{h}} = h_0 \hat{z}$ breaks $O(3) \to O(2)$. The **longitudinal mode** (fluctuations of $|\mathbf{h}|$) is gapped. The two **transverse modes** (perpendicular fluctuations) are gapless by the Goldstone theorem.

The transverse Hessian has a *first-order* time derivative $K_\perp \sim \chi_\perp \mathrm{i}\omega_n + \rho_s \mathbf{q}^2$ (reflecting spin precession), giving **quadratic** dispersion $\omega \propto q^2$. This contrasts with the *linear* $\omega \propto q$ of superfluids, where the time derivative is second-order. The spin waves are Goldstone modes of the broken $O(3)$ symmetry, but their quadratic dispersion reflects the precessional character of spin dynamics.

Deriving the magnon Hessian

The quadratic dispersion $\omega \propto q^2$ of spin waves can be traced to the structure of the fermionic $\mathrm{Tr}\log$ in the spin channel. The transverse collective-field Hessian is

$$K_\perp(\mathrm{i}\omega_n, \mathbf{q}) = \frac{1}{U} - \Pi_\perp(\mathrm{i}\omega_n, \mathbf{q}),$$

where $\Pi_\perp$ is the transverse (particle-hole) polarization bubble. For the Hubbard model with exchange splitting h_0, the transverse bubble involves propagators for up-spin particles and down-spin holes:

$$\Pi_\perp(\mathrm{i}\omega_n, \mathbf{q}) = -T \sum_{\nu_m} \int_{\mathbf{k}} G_\uparrow(\mathrm{i}\nu_m, \mathbf{k})\, G_\downarrow(\mathrm{i}\nu_m + \mathrm{i}\omega_n, \mathbf{k}+\mathbf{q}),$$

where $G_\sigma(\mathrm{i}\nu_m, \mathbf{k}) = 1/(\mathrm{i}\nu_m - \xi_{\mathbf{k}} + \sigma h_0)$ with $\sigma = \pm 1$ for up/down spins. Performing the Matsubara sum:

$$\Pi_\perp(\mathrm{i}\omega_n, \mathbf{q}) = \int_{\mathbf{k}} \frac{n_F(\xi_{\mathbf{k}} - h_0) - n_F(\xi_{\mathbf{k}+\mathbf{q}} + h_0)}{\mathrm{i}\omega_n - \xi_{\mathbf{k}+\mathbf{q}} - h_0 + \xi_{\mathbf{k}} - h_0}.$$

The key point is that the energy denominator is $\mathrm{i}\omega_n - (\xi_{\mathbf{k}+\mathbf{q}} + h_0) + (\xi_{\mathbf{k}} - h_0) = \mathrm{i}\omega_n - 2h_0 - (\xi_{\mathbf{k}+\mathbf{q}} - \xi_{\mathbf{k}})$. The shift $2h_0$ makes $\Pi_\perp$ depend on ω_n at zeroth order (not just through ω_n^2). Expanding at small

ω_n and q:

$$\Pi_\perp(\mathrm{i}\omega_n, \mathbf{q}) \approx \Pi_\perp(0,0) + \frac{\partial \Pi_\perp}{\partial(\mathrm{i}\omega_n)} \cdot \mathrm{i}\omega_n + \frac{1}{2}\frac{\partial^2 \Pi_\perp}{\partial q_i \partial q_j} q_i q_j + \cdots$$

The linear-in-ω_n term is the crucial difference from the superfluid case. It arises because spin precession is a first-order-in-time process (the Bloch equation $\dot{\mathbf{S}} = \mathbf{S} \times \mathbf{h}$ is first-order), unlike the density oscillation that governs superfluid sound (which is second-order, $\ddot{n} \propto \nabla^2 P$).

Using the Stoner condition $1/U = \Pi_\perp(0,0)$ (which ensures the saddle is self-consistent), the Hessian becomes

$$K_\perp(\mathrm{i}\omega_n, \mathbf{q}) = -\chi_\perp \mathrm{i}\omega_n + \rho_s q^2 + \cdots,$$

where $\chi_\perp = -\partial\Pi_\perp/\partial(\mathrm{i}\omega_n)|_0$ and ρ_s comes from the q^2 expansion. The mode condition $K_\perp = 0$ gives $\omega = -\mathrm{i}\rho_s q^2/\chi_\perp$—after analytic continuation $\mathrm{i}\omega_n \to \omega + \mathrm{i}0^+$, this becomes $\omega = (\rho_s/\chi_\perp)q^2$: the quadratic magnon dispersion.

The factor $\chi_\perp$ is the transverse susceptibility, and ρ_s is the spin stiffness. Both can be computed from the microscopic propagators. The ratio $\rho_s/\chi_\perp$ is the spin-wave velocity squared (for linear dispersion) or the diffusion-like coefficient (for quadratic dispersion). The first-order time derivative is the mathematical signature of spin precession, and its absence (replaced by a second-order derivative) is the mathematical signature of density oscillation.

Part IX

Beyond Gaussian Fluctuations

The RPA—the Gaussian approximation for collective-field fluctuations—is often remarkably successful, but it is not exact. This part develops the structure of corrections beyond the Gaussian level.

Chapter 35

Higher Derivatives of the Effective Action

The Taylor expansion of $S_{\text{eff}}[\bar{\sigma}+\eta]$ around the saddle contains terms of all orders:

$$S_{\text{eff}}[\bar{\sigma}+\eta] = S_{\text{eff}}[\bar{\sigma}] + \frac{1}{2}\eta K \eta + \frac{1}{3!}\Gamma^{(3)}\eta^3 + \frac{1}{4!}\Gamma^{(4)}\eta^4 + \cdots,$$

where $\Gamma^{(n)} = \frac{\delta^n S_{\text{eff}}}{\delta \sigma^n}|_{\bar{\sigma}}$ are the **effective vertices**. For $M(\sigma) = A_0 + \sigma\Gamma$ (linear dependence), the n-th derivative of $\operatorname{Tr}\log M$ gives cyclic traces:

$$\frac{\delta^n}{\delta\sigma(x_1)\cdots\delta\sigma(x_n)} \operatorname{Tr}\log M = (-1)^{n+1}(n-1)!\ \operatorname{tr}[G(x_1,x_2)\Gamma\cdots G(x_n,x_1)\Gamma]\,.$$

The derivation proceeds by induction: each differentiation introduces one more $-G\Gamma G$ factor via the identity $\delta G = -G(\delta M)G$. The resulting diagrams are closed polygons: tadpole ($n=1$), bubble ($n=2$), triangle ($n=3$), box ($n=4$), and so on.

The cubic vertex $\Gamma^{(3)}$ allows one collective mode to decay into two. The quartic vertex $\Gamma^{(4)}$ governs two-mode scattering. Within the Gaussian approximation, these are neglected; including them produces corrections to D, Σ, and the gap equation.

Chapter 36

Vertex Corrections

In the RPA, the polarization bubble $\Pi = \text{tr}[G\Gamma G\Gamma]$ is the Gaussian (Wick-factorized) approximation to the covariance of the composite observable O. The exact susceptibility includes **vertex corrections**:

$$\chi = \text{tr}[G\Lambda G\Gamma] \neq \text{tr}[G\Gamma G\Gamma],$$

where Λ is the dressed vertex function. In statistical language: vertex corrections arise because Wick factorization is not exact for non-Gaussian distributions. They measure residual correlations beyond the pairwise covariance—the connected four-point cumulant that the Gaussian approximation sets to zero.

Vertex corrections matter near phase transitions (where the Gaussian approximation breaks down), when exact identities are at stake (Ward identities), and at strong coupling. The vertex function Λ satisfies the **Bethe–Salpeter equation**, which resums vertex insertions just as the Dyson equation resums self-energy insertions.

36.1 The Bethe–Salpeter Equation

The Bethe–Salpeter equation is the vertex analogue of the Dyson equation. Just as the Dyson equation $G = G_0 + G_0\Sigma G$ resums self-energy insertions to all orders, the Bethe–Salpeter equation resums vertex insertions.

Define the **full vertex** $\Lambda(k, k+q)$ as the exact coupling between the composite observable and the microscopic field (generalizing the bare vertex Γ). The exact susceptibility is

$$\chi(q) = -\int_k \text{tr}[G(k+q)\Lambda(k, k+q)\, G(k)\Gamma].$$

Comparing with the RPA bubble $\Pi(q) = -\int_k \text{tr}[G(k+q)\Gamma\, G(k)\Gamma]$, the vertex correction replaces one bare vertex Γ by the full vertex Λ.

The Bethe–Salpeter equation for Λ is

$$\boxed{\Lambda(k, k+q) = \Gamma + \int_{k'} \mathcal{K}(k, k'; q)\, G(k'+q)\Lambda(k', k'+q)\, G(k'),} \tag{36.1}$$

where $\mathcal{K}(k, k'; q)$ is the **irreducible interaction kernel**—the analogue of the self-energy Σ for the two-particle channel. The equation has the structure of a Fredholm integral equation: the full vertex Λ equals the bare vertex Γ plus a convolution of Λ with the kernel $\mathcal{K}$ and two propagators.

In operator notation, writing $\hat{\Pi} = GG$ (the product of two propagators, viewed as an operator) and $\hat{\mathcal{K}}$ for the kernel:

$$\Lambda = \Gamma + \hat{\mathcal{K}}\hat{\Pi}\Lambda = (1 - \hat{\mathcal{K}}\hat{\Pi})^{-1}\Gamma.$$

This is a Neumann series: $\Lambda = \Gamma + \hat{\mathcal{K}}\hat{\Pi}\Gamma + (\hat{\mathcal{K}}\hat{\Pi})^2\Gamma + \cdots$. Each term corresponds to one more rung in a "ladder diagram." The susceptibility then becomes

$$\chi = -\operatorname{tr}[\hat{\Pi}\Lambda\Gamma] = -\operatorname{tr}[\hat{\Pi}(1 - \hat{\mathcal{K}}\hat{\Pi})^{-1}\Gamma^2].$$

The Bethe–Salpeter equation is to the vertex what the Dyson equation is to the propagator: an exact resummation in terms of an irreducible building block. In the RPA, $\Lambda = \Gamma$ (the vertex correction is neglected). The simplest correction beyond RPA is the ladder approximation $\mathcal{K} = V$ (the bare interaction), which produces a geometric series of ladder diagrams. Near a phase transition, the ladder sum diverges (the susceptibility diverges), signaling the onset of long-range order—exactly as the divergence of $D = (D_0^{-1} + \Pi)^{-1}$ at $D_0^{-1} + \Pi(0) = 0$ signals the phase transition in the collective channel.

Chapter 37

Self-Consistency and Conserving Approximations

Several quantities are interconnected: the gap equation involves the propagator, the propagator involves the self-energy, the self-energy involves the polarization, and the polarization involves the propagator and vertex. Changing one without adjusting the others can violate exact identities.

37.1 The Luttinger–Ward Functional

The **Baym–Kadanoff framework** resolves this by deriving all approximations from a single scalar functional $\Phi[G]$, called the **Luttinger–Ward functional**. This functional depends on the full propagator G (not on the bare parameters), and all physical quantities are obtained from it by differentiation:

- The self-energy: $\Sigma[G] = \delta\Phi/\delta G$.
- The propagator: determined self-consistently from $G = (G_0^{-1} + \Sigma[G])^{-1}$.
- The thermodynamic potential: $\Omega = \Phi[G] - \mathrm{Tr}[\Sigma G] + \mathrm{Tr}\log G$.

The crucial point is that $\Phi[G]$ depends on the full propagator G, not on the bare coupling g or the auxiliary-field mass $1/g$ from the HST. All reference to the "prior" (the bare action $S_0[\sigma] = \sigma^2/(2g)$) has been absorbed: the Luttinger–Ward functional is a functional of the physical observable G alone. This is a powerful conceptual simplification: the gap equation, the Hessian, and the collective propagator can all be expressed without reference to the channel in which the HST was performed.

Because all quantities derive from Φ by differentiation, Ward identities and conservation laws are automatically satisfied. An approximation constructed this way is called **conserving** or **Φ-derivable**.

37.2 Examples

The simplest conserving approximation is Hartree–Fock, where $\Phi[G]$ is the first-order self-energy diagram. The next level (including the $\mathrm{Tr}\log$ of the bubble) gives the GW approximation. More elaborate choices of Φ include vertex corrections and produce more accurate self-consistent equations.

Worked example: the Hartree approximation as a Φ-derivable theory

The simplest conserving approximation takes $\Phi[G]$ to be the first-order (Hartree) diagram:

$$\Phi_{\mathrm{H}}[G] = \frac{\lambda}{4}\left[\int_x G(x,x)\right]^2 = \frac{\lambda}{4}\left[\int_k G(k)\right]^2.$$

The self-energy is

$$\Sigma_{\mathrm{H}} = \frac{\delta\Phi_{\mathrm{H}}}{\delta G} = \frac{\lambda}{2}\int_k G(k) = \frac{\lambda}{2}G(x,x).$$

This is a constant (momentum-independent), and the self-consistent equation $G = (G_0^{-1} + \Sigma_{\mathrm{H}})^{-1}$ with $G_0^{-1}(k) = k^2 + m_0^2$ gives $G(k) = 1/(k^2+m^2)$ with $m^2 = m_0^2 + \frac{\lambda}{2}\int_k 1/(k^2+m^2)$. This is exactly the gap equation from the HST saddle (Chapter 13), confirming that the Hartree approximation equals the saddle-point approximation in the density channel.

The thermodynamic potential is $\Omega = \Phi_{\mathrm{H}}[G] - \frac{1}{2}\operatorname{Tr}[\Sigma_{\mathrm{H}}G] + \frac{1}{2}\operatorname{Tr}\log G = \frac{\lambda}{4}(\int_k G)^2 - \frac{\lambda}{4}(\int_k G)^2 + \frac{1}{2}\operatorname{Tr}\log G = \frac{1}{2}\operatorname{Tr}\log G$, which is the entropy contribution. All conservation laws (particle number, energy, momentum) are automatically satisfied because the approximation derives from a single functional.

The next level, $\Phi_{\mathrm{GW}}[G] = \Phi_{\mathrm{H}}[G] + \frac{1}{2}\operatorname{Tr}\log(1+G_0^{-1}\Sigma_{\mathrm{bubble}})$ (the "GW approximation" of Hedin), includes the bubble resummation and gives a momentum-dependent self-energy. This is the Φ-derivable version of the RPA.

The reader should view the Baym–Kadanoff framework not as a specific approximation, but as a *principle* for designing approximations that maintain internal consistency.

37.3 $S_{\text{eff}}[\sigma]$ vs. $F[G]$: Two Formulations

The auxiliary-field and Luttinger–Ward formulations are compared in the following table:

	$S_{\text{eff}}[\sigma]$ (auxiliary field)	$F[G]$ (Luttinger–Ward)
Variable	σ (auxiliary, non-physical)	G (propagator, physical)
Prior mass	Explicit: $\sigma^2/(2g)$	Absent (absorbed into Φ)
Stationarity	Gap eq.: $\bar{\sigma} = -gG_{nn}$	Dyson: $G^{-1} = K_0 + \Sigma[G]$
Tr log	$\frac{1}{2}\operatorname{Tr}\log K[\sigma]$	$-\frac{1}{2}\operatorname{Tr}\log G$ (entropy)
Conservation	Must check separately	Automatic (Baym–Kadanoff)
Algorithm	EM: iterate σ	Dyson iteration: iterate G
Equivalence	Same fixed point. Same G at convergence.	

The key insight is that at the saddle, $\bar{\sigma} = -gG_{nn}$, so the prior $\bar{\sigma}^2/(2g) = gG_{nn}^2/2$ is a *known function of* G. The auxiliary field σ and its prior are scaffolding: they help derive $F[G]$, but once F is written down, they can be discarded. The Luttinger–Ward functional $F[G] = \Phi[G] - \frac{1}{2}\operatorname{Tr}\log G$ depends on the physical observable G alone.

Part X

Symmetry, Ward Identities, and Local Geometry

Chapter 38

Ward Identities from Symmetry

38.1 Symmetry Implies Identities

If $S[T_\alpha \phi] = S[\phi]$ for a continuous family of transformations T_α, differentiating with respect to α produces identities among correlation functions. The mechanism is always the same: differentiate the invariance, take expectations, and read off the identity.

38.2 First Derivative: Conservation

Differentiate $S[T_\alpha \phi] = S[\phi]$ with respect to α at $\alpha = 0$, defining the infinitesimal generator $\delta_\alpha \phi(x) = \frac{d}{d\alpha}|_{\alpha=0}[T_\alpha \phi](x)$. This gives

$$\int \frac{\delta S}{\delta \phi(x)}\, \delta_\alpha \phi(x)\, \mathrm{d}x = 0$$

for all field configurations. Inside the path integral, this becomes $\langle \int \frac{\delta S}{\delta \phi} \delta_\alpha \phi \ \mathrm{d}x \rangle = 0$—a conservation equation relating expectation values.

38.3 Second Derivative: The Ward–Takahashi Identity

The second derivative produces the **Ward–Takahashi identity**:

$$q_\mu \Gamma^\mu(k, k{+}q) = G^{-1}(k{+}q) - G^{-1}(k), \tag{38.1}$$

which constrains the relationship between the self-energy (which modifies G^{-1}) and the vertex Γ^μ (which couples the field to the symmetry current). This is not optional: it is a mathematical consequence of the symmetry.

An approximation that modifies Σ without adjusting Γ^μ will violate (38.1) and, consequently, the conservation law. This is why conserving approximations (Chapter 32) matter in practice: a more sophisticated approximation that violates a Ward identity can give worse results than a simpler one that respects it.

38.4 Consistency Between Gap Equations and Zero Modes

The gap equation (first derivative of S_{eff}) and the Goldstone condition $K(0, 0)v_{\mathrm{sym}} = 0$ (second derivative of S_{eff}) must be mutually consistent. If computed at different levels of approximation, the Hessian may fail to have a zero eigenvalue where symmetry requires one—producing a spurious gap in a mode that should be gapless.

Worked example: Ward identity for the $O(N)$ model

Consider the $O(N)$ model from Chapter 29, with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ and action invariant under $O(N)$ rotations. In the broken phase with $\langle \boldsymbol{\phi} \rangle = (v, 0, \ldots, 0)$, the fields decompose into one longitudinal mode $\sigma = \phi_1 - v$ and $N - 1$ transverse (Goldstone) modes $\pi_a = \phi_a$ for $a = 2, \ldots, N$.

The $O(N)$ symmetry relates the longitudinal and transverse self-energies. An infinitesimal rotation in the $(1, a)$ plane, $\phi_1 \to \phi_1 - \epsilon\phi_a$, $\phi_a \to \phi_a + \epsilon\phi_1$, leaves the action invariant. Differentiating the generating functional $W[J]$ with respect to ϵ and then taking appropriate J-derivatives, we obtain the Ward identity:

$$G_\sigma^{-1}(q = 0) - G_\pi^{-1}(q = 0) = v \cdot \Gamma^{(3)}_{\sigma\pi\pi}(0, 0, 0) \cdot G_\pi(0), \tag{38.2}$$

where $\Gamma^{(3)}_{\sigma\pi\pi}$ is the three-point vertex coupling one σ to two π modes.

At the exact level, the Goldstone theorem requires $G_\pi^{-1}(0) = 0$ (the transverse mode is gapless). The Ward identity then constrains $G_\sigma^{-1}(0) = v \cdot \Gamma^{(3)}(0) \cdot G_\pi(0)$. Since $G_\pi(0)$ diverges (gapless mode), the product $\Gamma^{(3)} \cdot G_\pi(0)$ must remain finite, which constrains $\Gamma^{(3)}$.

The practical consequence is this: any approximation that computes G_σ and G_π independently (e.g., using different self-energy diagrams for the longitudinal and transverse channels) may violate (38.2). The Φ-derivable approximation of Chapter 33 automatically satisfies it, because both G_σ and G_π derive from the same functional $\Phi[G]$.

This example illustrates a general principle: Ward identities are constraints between different correlation functions that follow from symmetry. They are automatically satisfied in the exact theory and in Φ-derivable approximations, but can be violated in ad hoc approximations.

Chapter 39

Geometry of the Effective Action

The Hessian spectrum of S_{eff} at the saddle determines the local geometry of the negative log-density landscape:

Hessian structure	**Interpretation**
Large positive eigenvalue	Stiff direction: small fluctuations
Small positive eigenvalue	Soft direction: large fluctuations
Zero eigenvalue	Flat (symmetry/Goldstone) direction
Negative eigenvalue	Unstable saddle direction
Vanishing eigenvalue at a transition	Critical softening, diverging susceptibility

This spectrum connects to the **Fisher information matrix** (the expected Hessian of the negative log-likelihood), to **natural gradients** (preconditioning by the inverse Fisher information), and to the geometry of optimization. Phase transitions are geometric events in this landscape: an eigenvalue of the Hessian crosses zero as a parameter varies, changing the qualitative structure of the distribution.

Part XI

Renormalization as Multiscale Marginalization

Chapter 40

The Lattice as the Starting Point

40.1 Why Begin on a Lattice?

Throughout this book we have emphasized the finite-dimensional picture before passing to the continuum. Renormalization is where this emphasis becomes essential. Every divergence in continuum field theory is an artifact of taking a continuum limit before the physical predictions have been extracted. If we stay on a lattice—or more generally, if we work with a finite ultraviolet cutoff—there are no divergences. The theory is a well-defined probability distribution on a finite-dimensional space.

Let us therefore begin with a scalar field ϕ_i defined on a d-dimensional cubic lattice with spacing a. The action is

$$S[\phi;\theta,a] = a^d \sum_i \left[\frac{1}{2} \sum_\mu \left(\frac{\phi_{i+\hat{\mu}} - \phi_i}{a} \right)^2 + \frac{1}{2} m_0^2 \phi_i^2 + \frac{\lambda_0}{4!} \phi_i^4 \right], \tag{40.1}$$

where $\theta = (m_0^2, \lambda_0)$ are the bare parameters and the sum runs over lattice sites i and lattice directions $\hat{\mu}$. This defines a probability distribution $p(\phi;\theta,a) \propto e^{-S[\phi;\theta,a]}$ on the lattice field configurations.

All the observables we care about—correlation functions, susceptibilities, collective-mode spectra—are well-defined expectations under this distribution. The lattice spacing a provides a natural ultraviolet cutoff: no Fourier mode with $|k| > \pi/a$ exists. The theory has no divergences.

40.2 Physical Observables at Wavelength λ

The quantities we wish to compute are observables at a physical wavelength $\lambda \gg a$—long-distance correlations, low-energy scattering, macroscopic response functions. Denote such an observable by $\mathcal{O}(\theta, a, \lambda)$, which depends on the bare parameters θ, the lattice spacing a, and the observation scale λ.

The central question of renormalization is:

> Can we choose the bare parameters $\theta(a)$ as a function of the lattice spacing so that the physical observable has a well-defined limit as $a \to 0$?

More precisely, we seek a function $\theta(a) = (m_0^2(a), \lambda_0(a))$ such that

$$\boxed{\mathcal{O}\big(\theta(a),\, a,\, \lambda\big) = \mathcal{O}_{\text{phys}} + o(a/\lambda)} \tag{40.2}$$

for all $\lambda > \lambda_{\min}$, where $\mathcal{O}_{\text{phys}}$ is the physical (continuum) value and $o(a/\lambda)$ denotes terms that vanish as $a/\lambda \to 0$. The lattice corrections are controlled by the ratio of the lattice spacing to the observation wavelength.

This is the definition of a **renormalized** theory: a family of lattice theories, parameterized by a, whose long-distance predictions converge to a common limit. The function $\theta(a)$ is the **bare-parameter trajectory** or **renormalization group trajectory**.

40.3 How Many Parameters Must Be Tuned?

Whether this program succeeds, and how many parameters must be tuned, depends on the dimension d and the structure of the action.

For the scalar ϕ^4 theory in $d < 4$ dimensions, the lattice theory (40.1) has two parameters (m_0^2, λ_0). As $a \to 0$, we need to tune $m_0^2(a)$ (which diverges logarithmically or as a power of $1/a$, depending on d) to keep the physical mass m_{phys} fixed. The coupling λ_0 may also need tuning in $d = 4$, but in $d < 4$ it flows to an infrared fixed point and requires no fine-tuning for the continuum limit. In $d > 4$, the coupling flows to zero (the Gaussian fixed point), and the continuum theory is free—interacting ϕ^4 theory does not exist in $d > 4$ (this is the **triviality** of ϕ^4 in high dimensions).

A theory is **renormalizable** if only a finite number of parameters need to be tuned as $a \to 0$. It is **non-renormalizable** (or **effective**) if the number of parameters that need tuning grows without bound. But even a non-renormalizable theory makes perfectly good predictions at wavelengths $\lambda \gg a$: the corrections are suppressed by powers of a/λ.

From the lattice-first perspective, we can reformulate renormalizability as **rendering stability**: a lattice theory "renders" stable predictions at long wavelengths, insensitive to sub-lattice structure, with only finitely many parameters to fix. The analogy is to computer graphics: the image at resolution λ is stable against changes in sub-pixel (sub-lattice) details, provided a finite number of rendering parameters (masses, couplings) are held fixed. The Symanzik effective theory makes this precise: lattice observables at $\lambda \gg a$ can be expanded as

$$\mathcal{O}(\theta, a, \lambda) = \mathcal{O}_{\text{cont}}(\lambda) + c_1 \left(\frac{a}{\lambda}\right)^p + c_2 \left(\frac{a}{\lambda}\right)^{p+1} + \cdots ,$$

where the leading correction is suppressed by a power p determined by the dimension of the first irrelevant operator.

Chapter 41

Dimensional Analysis and the Classification of Interactions

The structure of the RG flow—which couplings are relevant, marginal, or irrelevant—is determined almost entirely by dimensional analysis. This chapter develops the dimensional-analysis framework systematically, both because it is essential for understanding renormalization and because it provides the deep reason for the UV–IR inversion discussed in the heat-kernel chapter.

41.1 Engineering Dimensions

In natural units ($\hbar = c = k_B = 1$), there is a single independent dimension, which we take to be mass (equivalently, inverse length). The action S is dimensionless (it appears in the exponent e^{-S}). From the kinetic term $\frac{1}{2}\int \mathrm{d}^d x\,(\nabla\phi)^2$, we read off the dimensions:

$$[x] = -1, \quad [\mathrm{d}^d x] = -d, \quad [\nabla] = +1, \quad [\phi] = \frac{d-2}{2}.$$

The field dimension $[\phi] = (d-2)/2$ is chosen to make the kinetic term dimensionless:

$$[\mathrm{d}^d x\,(\nabla\phi)^2] = -d + 2 + 2\cdot\frac{d-2}{2} = 0. \qquad \checkmark$$

With this convention, the dimensions of the couplings in $S = \int \mathrm{d}^d x\left[\frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m^2\phi^2 + \frac{\lambda}{4!}\phi^4 + \frac{g_6}{6!}\phi^6 + \cdots\right]$ are:

Operator	**Symbol**	**Dimension of coupling**	**In $d = 4$**
$(\nabla\phi)^2$	(kinetic)	0 (normalization)	0
ϕ^2	m^2	$+2$	$+2$
ϕ^4	λ	$4-d$	0
ϕ^6	g_6	$6-2d$	-2
ϕ^{2n}	g_{2n}	$n(d-2)-d$	$2n-4-n(d-2)$
$\phi^2(\nabla\phi)^2$	h_4	$2-d$	-2
$(\nabla\phi)^4$	h_4'	$4-d$	0 (marginal)

The **engineering dimension** of a coupling determines its behavior under the rescaling step of the RG:

$$g \to g' = b^{[g]} \cdot g,$$

where $b > 1$ is the rescaling factor. If $[g] > 0$, the coupling grows under coarse-graining: it is **relevant**. If $[g] = 0$, it is **marginal**. If $[g] < 0$, it shrinks: it is **irrelevant**.

This classification has a direct physical meaning:

Relevant couplings ($[g] > 0$) control the long-distance physics. They must be tuned (measured) to make predictions. Each relevant coupling corresponds to one free parameter of the theory. Examples: the mass m^2 (always relevant for $d > 0$), the cosmological constant.

Marginal couplings ($[g] = 0$) are the borderline cases. Their fate is determined by the sign of the interaction correction (the "beta function"): if the correction makes them grow, they are "marginally relevant"; if it makes them shrink, "marginally irrelevant." Example: ϕ^4 coupling in $d = 4$.

Irrelevant couplings ($[g] < 0$) die away under coarse-graining. Their effects are suppressed by powers of a/λ at long distances. They can be set to any value without affecting the long-distance predictions—this is the origin of universality. Example: ϕ^6 coupling in $d \leq 3$.

41.2 Why Dimensional Analysis Controls the RG

The dominance of dimensional analysis over the RG flow is not an approximation—it is exact in the following precise sense. Near a *free* (Gaussian) fixed point, the RG eigenvalues (scaling dimensions) equal the engineering dimensions plus corrections of order λ^*. Since the Gaussian fixed point has $\lambda^* = 0$, the corrections vanish, and the engineering dimensions are exact.

Near an *interacting* fixed point (like Wilson–Fisher), the corrections are nonzero but typically small: the anomalous dimensions γ satisfy $|\gamma| \ll 1$ for the operators we consider. The scaling dimension of an operator is $\Delta = \Delta_{\text{eng}} + \gamma$, where Δ_{eng} is the engineering ("naive") dimension and γ is the anomalous correction. The classification into relevant/marginal/irrelevant changes only when γ is large enough to cross the marginal threshold $\Delta = d$—which happens rarely.

The deep reason that dimensional analysis controls the classification is that the RG is a *semigroup of rescalings*, and the engineering dimensions are the eigenvalues of the linearized rescaling at the free fixed point. Interactions perturb these eigenvalues but do not (generically) change their signs.

41.3 Connection to the UV–IR Inversion

The UV–IR inversion of the heat-kernel chapter is now seen as a corollary of dimensional analysis.

An operator $\mathcal{O}_n$ with engineering dimension $\Delta_n = n(d-2)/2 + 2n - d$ contributes to the effective action with a coefficient that scales as $\Lambda^{d-\Delta_n}$. This is the UV sensitivity. The same operator contributes to long-distance physics with a weight $k^{\Delta_n - d}$ relative to the leading operator. The UV sensitivity $\Lambda^{d-\Delta_n}$ and the IR weight k^{Δ_n-d} are *reciprocal powers of the same exponent* $d - \Delta_n$: one measures how fast the coefficient diverges at short distances, the other measures how fast the contribution grows at long distances.

The operator with $\Delta_n = 0$ (the cosmological constant, or the identity operator) has UV sensitivity Λ^d and IR weight k^0: most UV-sensitive, most IR-relevant. The operator with $\Delta_n = d$ (a marginal operator) has UV sensitivity $\log \Lambda$ and IR weight $k^0 \cdot \log(k)$: logarithmic in both. The operator with $\Delta_n > d$ (irrelevant) has UV sensitivity $\Lambda^{d-\Delta_n} \to 0$ and IR weight $k^{\Delta_n-d} \to 0$: UV-finite and IR-suppressed.

41.4 Counting Free Parameters: Relevant and Marginal Operators

The number of free parameters in the continuum theory equals the number of relevant and marginal operators at the fixed point. This is the precise answer to the question “how many measurements are needed to determine the theory?”:

For the ϕ^4 theory in $d = 4$: two relevant (m^2, the vacuum energy ρ_Λ) and one marginal (λ) $\Rightarrow$ three free parameters.

For the ϕ^4 theory in $d = 3$: two relevant (m^2, ρ_Λ) and λ is now relevant too $\Rightarrow$ three free parameters (but λ flows to the Wilson–Fisher fixed point, so at criticality one parameter is fixed by universality).

For the $O(N)$ model: same counting, with $\lambda \to \lambda/N$.

For general relativity in $d = 4$: two relevant (ρ_Λ, Newton's constant G) and finitely many marginal and irrelevant higher-derivative terms. The theory is non-renormalizable in the traditional sense, but it has a finite number of relevant couplings and makes predictions at wavelengths $\lambda \gg \ell_{\text{Planck}}$.

Chapter 42

Coarse-Graining and the Wilsonian RG for Scalar Fields

42.1 Splitting the Field

Following Wilson, we split the lattice field into slow (long-wavelength) and fast (short-wavelength) components:

$$\phi = \phi_< + \phi_>,$$

where $\phi_<$ contains Fourier modes with $|k| < \Lambda/b$ (for some scaling factor $b > 1$) and $\phi_>$ contains modes with $\Lambda/b < |k| < \Lambda$, where $\Lambda = \pi/a$ is the lattice cutoff.

The effective action for the slow modes is obtained by marginalization:

$$e^{-S_{\text{eff}}[\phi_<]} = \int \mathcal{D}\phi_> \, e^{-S[\phi_< + \phi_>]}. \tag{42.1}$$

This is exactly the marginalization operation we have used throughout the book—integrating out latent (fast) variables to obtain an effective description of the remaining (slow) variables.

42.2 The Cumulant Expansion for ϕ^4

For the scalar ϕ^4 theory, write $S[\phi_< + \phi_>] = S_0[\phi_<] + S_0[\phi_>] + S_{\text{int}}[\phi_<, \phi_>]$, where S_0 is the free (quadratic) action and S_{int} contains the quartic coupling. The interaction couples slow and fast modes:

$$S_{\text{int}} = \frac{\lambda_0}{4!} \int (\phi_< + \phi_>)^4 = \frac{\lambda_0}{4!} \int \left[\phi_<^4 + 4\phi_<^3\phi_> + 6\phi_<^2\phi_>^2 + 4\phi_<\phi_>^3 + \phi_>^4\right].$$

Integrating out $\phi_>$ perturbatively using the cumulant expansion:

$$-\log\langle e^{-S_{\text{int}}}\rangle_> = \langle S_{\text{int}}\rangle_> - \frac{1}{2}\big(\langle S_{\text{int}}^2\rangle_> - \langle S_{\text{int}}\rangle_>^2\big) + \cdots, \tag{42.2}$$

where $\langle\cdot\rangle_>$ denotes expectation over the fast modes with the free action.

First cumulant (tadpole correction). The term $\langle 6\phi_<^2\phi_>^2\rangle_> = 6\phi_<^2\langle\phi_>^2\rangle_>$, where $\langle\phi_>(x)^2\rangle_> = \int_{\Lambda/b}^{\Lambda} \frac{d^dk}{(2\pi)^d}\frac{1}{k^2+m_0^2}$ is the variance of the fast modes. This produces a correction to the mass:

$$\delta m^2 = \frac{\lambda_0}{2}\int_{\Lambda/b}^{\Lambda} \frac{d^dk}{(2\pi)^d}\frac{1}{k^2 + m_0^2}. \tag{42.3}$$

This is the tadpole contribution we derived in Chapter 28, now restricted to the fast-mode shell.

Second cumulant (bubble correction). The connected part of $\langle S_{\text{int}}^2 \rangle_>$ produces corrections to the quartic coupling and generates new six-point interactions. The leading correction to λ_0 comes from the bubble of fast-mode propagators:

$$\delta\lambda = -\frac{3\lambda_0^2}{2} \int_{\Lambda/b}^{\Lambda} \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2 + m_0^2)^2}. \tag{42.4}$$

This is negative: integrating out fast modes *reduces* the effective coupling (in $d < 4$).

42.3 Rescaling

After integrating out the shell $\Lambda/b < |k| < \Lambda$, the remaining field $\phi_<$ has cutoff Λ/b. To compare with the original theory, we rescale:

$$x \to x' = x/b, \qquad k \to k' = bk, \qquad \phi_<(x) \to \phi'(x') = b^{(d-2)/2} \phi_<(bx'),$$

where the field rescaling exponent $(d-2)/2$ is chosen to preserve the kinetic term $\frac{1}{2}(\nabla\phi)^2$. Under rescaling:

$$m^2 \to m'^2 = b^2(m^2 + \delta m^2), \qquad \lambda \to \lambda' = b^{4-d}(\lambda + \delta\lambda).$$

The factors b^2 and b^{4-d} are the **engineering dimensions**—the scaling that each parameter would have in the absence of interactions.

42.4 The RG Flow Equations

For an infinitesimal step $b = e^\ell \approx 1 + \ell$, the RG flow equations are (to leading order in λ):

$$\frac{dm^2}{d\ell} = 2m^2 + c_d \frac{\lambda\Lambda^{d-2}}{m^2 + \Lambda^2}, \tag{42.5}$$

$$\frac{d\lambda}{d\ell} = (4-d)\lambda - c_d' \frac{3\lambda^2}{(m^2 + \Lambda^2)^2} \Lambda^{d-4}, \tag{42.6}$$

where c_d and c_d' are dimension-dependent constants from the angular integrals. In $d = 4 - \epsilon$ with the standard normalization, these are

$$c_d = \frac{S_d \Lambda^d}{(2\pi)^d} \cdot \frac{1}{(\Lambda^2 + m^2)}, \qquad c_d' = \frac{S_d \Lambda^d}{(2\pi)^d} \cdot \frac{1}{(\Lambda^2 + m^2)^2},$$

where $S_d = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of the unit sphere ($S_4 = 2\pi^2$). Defining the dimensionless coupling $\bar{\lambda} = \lambda\Lambda^{d-4} S_d/[(2\pi)^d]$ and working at the critical point $m^2 = 0$ in $d = 4$:

$$\frac{d\bar{\lambda}}{d\ell} = \epsilon\,\bar{\lambda} - \frac{3\bar{\lambda}^2}{2} + O(\bar{\lambda}^3).$$

The numerical coefficient 3/2 is specific to a single real scalar field; for the $O(N)$ model it becomes $(N+8)/6$.

Let us read the physics from these equations.

The mass has engineering dimension +2 (the $2m^2$ term), meaning it grows under coarse-graining. Near the Gaussian fixed point, m^2 is a **relevant** parameter: small changes in the bare mass produce large changes in the long-distance physics. This is why the mass must be tuned as $a \to 0$.

The coupling λ has engineering dimension $4-d$. In $d < 4$, the first term $(4-d)\lambda > 0$ drives λ to grow—λ is relevant, and the theory flows to an interacting fixed point (the Wilson–Fisher fixed point). In $d = 4$, the engineering dimension vanishes, λ is **marginal**, and the flow is controlled by the second term: $d\lambda/d\ell \approx -c\lambda^2 < 0$, so λ decreases logarithmically under coarse-graining. In $d > 4$, λ is **irrelevant**: it shrinks under coarse-graining, and the long-distance theory is Gaussian.

42.5 Generated Operators

The shell integration generates operators not present in the original action. At second order in the cumulant expansion, the contraction $\langle \phi_<^2 \phi_>^2 \cdot \phi_<^2 \phi_>^2 \rangle_>^c$ produces (among other terms) a sixth-order coupling:

$$\delta S_6 = \frac{\lambda^2}{6} \int \phi_<^6 \cdot \langle \phi_>(x)^2 \phi_>(y)^2 \rangle_>^c + \cdots$$

This generates a ϕ^6 interaction that was absent from the initial action. Its coefficient is $O(\lambda^2)$, and under rescaling it acquires the engineering dimension $[g_6] = 6 - 2d$. In $d = 3$, $[g_6] = 0$ (marginal at engineering level), but the generated value is $O(\lambda^2) \sim O(\epsilon^2)$ and the anomalous dimension pushes it to be irrelevant near the Wilson–Fisher fixed point. In $d = 4 - \epsilon$, $[g_6] = 2 - 2\epsilon < 0$ for small ϵ, so g_6 is irrelevant.

Similarly, derivative interactions $(\nabla\phi)^4$, $\phi^2(\nabla\phi)^2$, and higher-order terms are generated. Their engineering dimensions ensure irrelevance near $d = 4$.

The truncation of the RG flow to finitely many couplings (m^2, λ) involves a **projection**: at each step, we discard the generated irrelevant operators and keep only the relevant and marginal ones. This projection is justified near a fixed point because the irrelevant operators flow to zero under iteration. Away from a fixed point, or in a strongly coupled theory, the projection may fail and the truncated flow may be unreliable.

42.6 Fixed Points

The Gaussian fixed point $m^2 = \lambda = 0$ exists in all dimensions. In $d < 4$, there is an additional **Wilson–Fisher fixed point** at $\lambda^* = (4-d)(m^2 + \Lambda^2)^2/(3c_d' \Lambda^{d-4})$, which controls the critical behavior of second-order phase transitions.

At a fixed point, the theory is scale-invariant: it looks the same at all length scales. The critical exponents (the power-law behavior of correlation functions at the critical point) are determined by the eigenvalues of the linearized RG flow around the fixed point.

42.7 The ϵ-Expansion and Critical Exponents

The most powerful tool for computing critical exponents is the ϵ-expansion, due to Wilson and Fisher. Setting $d = 4 - \epsilon$ with ϵ small, the coupling λ is weakly relevant (its engineering dimension is ϵ), and the Wilson–Fisher fixed point is perturbatively accessible.

The fixed-point coupling

From the flow equation (42.6), setting $d\lambda/d\ell = 0$:

$$0 = \epsilon\lambda^* - c\lambda^{*2},$$

where $c = 3/(16\pi^2)$ in $d = 4$ (the coefficient of the bubble integral). This gives the fixed-point coupling

$$\lambda^* = \frac{\epsilon}{c} = \frac{16\pi^2\epsilon}{3}.$$

The fixed point is at a finite, computable value of the coupling, proportional to ϵ. This justifies the perturbative approach: at small ϵ, the fixed point is weakly coupled.

Critical exponents from linearized RG

At the fixed point, linearize the RG flow. The Jacobian matrix $\partial(\dot{m}^2, \dot{\lambda})/\partial(m^2, \lambda)|_{m^2=0,\lambda=\lambda^*}$ has eigenvalues that determine the critical exponents. The relevant eigenvalue is

$$y_t = 2 - \frac{\lambda^*}{c'} \cdot \frac{\partial}{\partial m^2}\left[\frac{c_d\lambda\Lambda^{d-2}}{m^2 + \Lambda^2}\right]_{m=0} = 2 - \frac{\epsilon}{3} + O(\epsilon^2),$$

where c' is a numerical coefficient. The correlation-length exponent is $\nu = 1/y_t$:

$$\boxed{\nu = \frac{1}{2} + \frac{\epsilon}{12} + O(\epsilon^2).} \tag{42.7}$$

At $\epsilon = 0$ (i.e., $d = 4$): $\nu = 1/2$, the mean-field (Gaussian) result. At $\epsilon = 1$ (i.e., $d = 3$): $\nu \approx 0.583$, compared to the exact value $\nu \approx 0.630$ for the 3D Ising universality class (the $N = 1$ case). Higher-order calculations in the ϵ-expansion, combined with resummation techniques, give $\nu \approx 0.628$—remarkably close to the exact answer from Monte Carlo simulations and the conformal bootstrap.

Anomalous dimensions

In the Gaussian theory, the two-point function at the critical point ($m = 0$) decays as $G(r) \sim 1/r^{d-2}$, which corresponds to the field having scaling dimension $[\phi] = (d-2)/2$. Interactions modify this to $[\phi] = (d - 2 + \eta)/2$, where η is the **anomalous dimension**. Physically, η measures the deviation of the critical-point correlator from the free-field form.

The anomalous dimension first appears at two-loop order in the ϵ-expansion:

$$\eta = \frac{(N+2)}{2(N+8)^2}\epsilon^2 + O(\epsilon^3).$$

For the Ising model ($N = 1$): $\eta = \frac{1}{54}\epsilon^2 \approx 0.019$ at $\epsilon = 1$, compared to the exact value $\eta \approx 0.036$. The smallness of η explains why the mean-field ($\eta = 0$) approximation works so well for many critical phenomena.

The anomalous dimension arises because the field rescaling in the RG step acquires a correction: instead of $\phi \to b^{(d-2)/2}\phi$ (the engineering scaling), the correct rescaling is $\phi \to b^{(d-2+\eta)/2}\phi$, reflecting the fact that fluctuation corrections modify the effective dimension of the field operator. In the language of this book, η is a correction to the exponent in the eigendecomposition of the covariance operator at the critical fixed point.

Chapter 43

Renormalization: Three Perspectives

43.1 The Wilsonian Perspective

Wilson's perspective, which we have just developed, views renormalization as a flow in the space of all possible actions. Starting from a lattice action with cutoff Λ, we integrate out momentum shells and track how the effective action changes. The key insight is that only a finite number of parameters (the relevant and marginal couplings) matter for long-distance physics. The irrelevant couplings die away under the flow, regardless of their initial values. This explains **universality**: different microscopic theories (with different irrelevant couplings) flow to the same fixed point and produce the same long-distance behavior.

43.2 The Weinberg Perspective: Asymptotic Safety and Effective Field Theory

Weinberg's perspective focuses on the question: for which theories does the continuum limit $a \to 0$ exist? In the language of equation (40.2), this asks: for which theories can we find a trajectory $\theta(a)$ along which the physical observables converge?

A theory is **perturbatively renormalizable** if the trajectory $\theta(a)$ can be parameterized by a finite number of physical quantities (masses, couplings measured at some reference scale). This is the case for ϕ^4 in $d = 4$, QED, and the Standard Model.

More broadly, Weinberg introduced the concept of **asymptotic safety**: a theory has a good continuum limit if its RG trajectory approaches a fixed point (not necessarily the Gaussian one) as $a \to 0$. The fixed point can be interacting, and the number of relevant directions at that fixed point determines how many parameters must be fixed by experiment.

Even if the continuum limit does not exist (as is likely for ϕ^4 in $d = 4$, which may be **trivial**), the theory makes perfectly good predictions at any finite cutoff. This is the **effective field theory** perspective: a theory with cutoff a and parameters θ makes predictions at wavelengths $\lambda > a$ with corrections of order $(a/\lambda)^p$ for some power p determined by the first irrelevant operator. The theory is valid as a low-energy approximation, even if it cannot be extended to arbitrarily short distances.

43.3 The Condensed-Matter Perspective: The Lattice Is Physical

In condensed-matter physics, the lattice is not a regularization device—it is the physical system. A crystal has atoms at discrete sites with spacing a. The question is not whether the continuum limit exists, but what the long-distance behavior is for a given lattice theory.

From this perspective, equation (40.2) reads differently: a is fixed (the physical lattice spacing), λ is the wavelength at which we observe, and the corrections $o(a/\lambda)$ are real lattice effects that can be measured experimentally. The bare parameters θ are determined by the microscopic physics (electronic band structure, phonon frequencies, interaction strengths), not by a renormalization prescription.

The universality of critical behavior—the fact that many different materials exhibit the same critical exponents near a phase transition—is explained by the RG: different microscopic Hamiltonians (different θ and even different lattice structures) flow to the same fixed point. The critical exponents are properties of the fixed point, not of the microscopic theory.

43.4 Unifying the Three Perspectives

Despite the different emphases, the three perspectives describe the same mathematics:

	What is a?	**What is the question?**
Wilson	Floating cutoff	How does S_{eff} change under coarse-graining?
Weinberg	Regulator $\to 0$	Does $\theta(a)$ exist with $\mathcal{O} \to \mathcal{O}_{\text{phys}}$?
Condensed matter	Physical lattice	What is the long-distance behavior?

In all three cases, the mathematical content is the same: marginalize over short-distance modes, track the effective parameters, and identify the long-distance predictions with the physical observables.

43.5 Renormalizability Is Not Required on the Lattice

The most important conceptual lesson of the lattice-first perspective is that **renormalizability is not a prerequisite for physical predictivity**. It is a property of the continuum limit $a \to 0$, not of the lattice theory at finite a.

The stability condition

Consider a lattice theory with spacing a, bare parameters $\theta(a)$, and an observable $\mathcal{O}$ measured at a physical wavelength $\lambda \geq \lambda_{\min} \gg a$. The theory makes reliable predictions whenever the **stability condition**

$$\boxed{\mathcal{O}(\theta(a),\, a,\, \lambda) = \mathcal{O}_{\text{phys}} + o(a/\lambda)} \tag{43.1}$$

is satisfied: the observable at wavelength λ depends on the lattice structure only through corrections that vanish as $a/\lambda \to 0$.

This condition says *nothing about UV completion or renormalizability.* It does not require the existence of a continuum limit $a \to 0$. It does not require the bare parameters $\theta(a)$ to be given by a perturbative counterterm expansion. It only requires that the lattice artifacts—the differences between the lattice prediction and the continuum prediction—be small at the observation scale.

What the stability condition requires

The stability condition (43.1) is satisfied whenever three ingredients are in place:

1. A lattice theory that is well-defined. The partition function $Z = \int \prod_i d\phi_i \, e^{-S[\phi;\theta,a]}$ converges. On the lattice, this is guaranteed by the finite number of integration variables and the positivity of e^{-S} (for real actions bounded below). No UV divergences exist.

2. A hierarchy of scales. The observation scale λ is much larger than the lattice spacing: $\lambda/a \gg 1$. This is the regime where the long-wavelength effective description applies.

3. Tuned bare parameters. The bare parameters $\theta(a)$ are chosen so that the relevant physical quantities (masses, coupling constants at scale $\lambda_{\min}$) take their desired values. The number of parameters that must be tuned equals the number of relevant directions at the nearest RG fixed point.

When these conditions hold, the Symanzik effective-theory analysis guarantees that all observables at scale λ deviate from their continuum values by at most $(a/\lambda)^p$ for some power $p \geq 1$ determined by the leading irrelevant operator:

$$\mathcal{O}(\theta(a), a, \lambda) = \mathcal{O}_{\text{cont}}(\lambda) + c_1 \left(\frac{a}{\lambda}\right)^p + c_2 \left(\frac{a}{\lambda}\right)^{p+1} + \cdots \tag{43.2}$$

The leading power p depends on the lattice action: for the standard nearest-neighbor action, $p = 2$ (the leading irrelevant operator has dimension $d+2$, contributing $(a/\lambda)^2$ corrections). For improved actions (Symanzik improvement), p can be increased by adding higher-order lattice operators tuned to cancel the leading corrections.

Non-renormalizable theories on the lattice

A "non-renormalizable" theory is one where the number of relevant and marginal couplings at the Gaussian fixed point is infinite, or where no non-trivial fixed point exists to which the RG trajectory can flow as $a \to 0$. In the continuum, this means the theory cannot be defined at arbitrarily short distances without introducing infinitely many free parameters.

But on the lattice, with $a > 0$ fixed, the theory is perfectly well-defined. The question is not "does the continuum limit exist?" but "does the theory make stable predictions at the observation scale λ?" And the answer is: *yes, provided* $\lambda \gg a$, with corrections suppressed by powers of a/λ.

The paradigmatic example is **general relativity**. The Einstein–Hilbert action is non-renormalizable by power counting: the graviton self-coupling has dimension $[G] = 2-d$, which is irrelevant in $d > 2$. In the continuum, this means that taking $a \to 0$ at fixed G produces uncontrolled divergences at each loop order.

But a lattice formulation of gravity (such as Regge calculus or causal dynamical triangulations) with fixed spacing a is a well-defined statistical mechanics problem. At wavelengths $\lambda \gg a$ (i.e., $\lambda \gg \ell_{\text{Planck}}$ if $a \sim \ell_{\text{Planck}}$), the predictions match general relativity with corrections of order $(\ell_{\text{Planck}}/\lambda)^2$. No UV completion is needed: the lattice *is* the UV completion.

The effective field theory of gravity makes quantitative predictions at $\lambda \gg \ell_{\text{Planck}}$—graviton scattering amplitudes, post-Newtonian corrections, gravitational wave templates—all as expansions in $(\ell_{\text{Planck}}/\lambda)^2$. These predictions are stable against changes in the UV completion (the lattice structure, string theory, loop quantum gravity, or any other Planck-scale physics), because the UV physics affects only the irrelevant operators, whose contributions are suppressed by $(a/\lambda)^p$.

What renormalizability does and does not guarantee

Let us be precise about the role of renormalizability:

Renormalizability guarantees: the existence of a continuum limit $a \to 0$ with finitely many free parameters. The theory can be defined at all length scales simultaneously. The predictions improve without bound as $a \to 0$.

Renormalizability does NOT guarantee: that the theory describes nature. A renormalizable theory is self-consistent as a mathematical structure, but it may still be an approximation to a more fundamental (possibly non-renormalizable) theory.

Non-renormalizability guarantees: that the continuum limit $a \to 0$ either does not exist or requires infinitely many free parameters. The theory cannot be defined at arbitrarily short distances with finite data.

Non-renormalizability does NOT prevent: stable, accurate, predictive physics at any finite ratio a/λ. A non-renormalizable theory with a physical lattice spacing a is as predictive as a renormalizable one, provided λ/a is large enough.

The stability condition (43.1) is the physically relevant criterion. Renormalizability is a *sufficient* condition for stability (it guarantees that $\theta(a)$ can be tuned with finitely many parameters for all a), but it is not *necessary*: any well-defined lattice theory with $\lambda/a \gg 1$ satisfies the stability condition, regardless of its renormalizability status.

The analogy to Taylor series

A useful analogy from the ML/statistics audience's experience: renormalizability is to field theory as convergence is to Taylor series.

A convergent Taylor series (like $e^x = 1+x+x^2/2+\cdots$) can be extended to arbitrarily large order and converges to a well-defined function. This is like a renormalizable theory: it can be extended to all energy scales.

An asymptotic Taylor series (like the Stirling series for $\log\Gamma(x)$) diverges at high order, but its first N terms provide excellent approximations when x is large. This is like a non-renormalizable effective theory: it breaks down at short distances, but it provides accurate predictions at long distances.

In practice, we almost always use Taylor series in the asymptotic regime (keeping a finite number of terms in a perturbative expansion), not in the convergent regime (summing the entire series). Similarly, in physics, we almost always use field theories as effective theories (at a finite ratio λ/a), not as exact continuum theories. The question of whether the series converges (whether the theory is renormalizable) is mathematically interesting but physically secondary.

Predictivity at finite a/λ: what matters in practice

For a reader of this book aiming to do research in physics, the practical message is:

Every lattice field theory is a well-defined probability model. No UV divergences, no regularization ambiguities, no need for renormalizability. The functional integrals, Dyson equations, gap equations, Hessians, and collective modes of this book are all defined on the lattice without any subtlety.

The lattice artifacts are small and controllable. For any observable at scale λ, the lattice corrections are of order $(a/\lambda)^p$, with $p \geq 2$ for standard actions and higher for improved actions. In practice, $a/\lambda \sim 0.01$–0.1 gives per-cent to per-mille accuracy.

The number of free parameters is finite. Even for a "non-renormalizable" theory, the number of relevant couplings at any given RG fixed point is finite. These are the parameters that must be determined by matching to experiment (or to a more fundamental theory). All other couplings are irrelevant and produce only $(a/\lambda)^p$ corrections.

UV completion is a question about short distances, not about long-distance predictions. Whether the theory has a good continuum limit ($a \to 0$), what happens at the Planck scale, or whether the theory is "fundamental"—these are fascinating questions, but they do not affect the predictions at $\lambda \gg a$. The stability condition (43.1) is the criterion that matters.

43.6 Divergences as Artifacts

From the lattice-first perspective, "ultraviolet divergences" are not physical. They are artifacts of taking the continuum limit $a \to 0$ before extracting physical predictions. On the lattice, every integral is finite. The tadpole integral $\int_k 1/(k^2 + m^2)$ is automatically cut off at $|k| = \pi/a$.

When we take $a \to 0$ with fixed bare parameters θ, the lattice integrals diverge because we are asking the wrong question: the bare parameters must change with a to keep the physics fixed. Renormalization is not about "removing infinities"—it is about correctly parameterizing the a-dependence of the bare theory so that physical predictions are a-independent (up to $o(a/\lambda)$ corrections).

The algebraic identities of this book—Dyson equations, Ward identities, saddle-point conditions, Hessian structures—hold on the lattice exactly as they do in the continuum. The lattice provides a concrete, finite-dimensional realization of every formula we have derived.

Part XII

Real-Time Structure, Spectra, and the Born Rule

Throughout this book, we have worked in the Euclidean formulation, treating the field ϕ as a random variable with density $p(\phi) \propto e^{-S_E[\phi]}$. This part develops the connection between the Euclidean statistical framework and real-time quantum mechanics, introducing the Lorentzian path integral, Wick rotation, spectral functions, and the physical interpretation of poles and widths. We pay particular attention to the question of when a quasiparticle description is valid and how quasiparticle modes differ from hydrodynamic modes.

Chapter 44

The Lorentzian Path Integral and Wick Rotation

44.1 Oscillatory vs. Gaussian: The Core Distinction

Before introducing path integrals, let us see the essential mechanism in a single integral. Consider two integrals that look superficially similar but are fundamentally different:

$$I_E = \int_{-\infty}^{\infty} e^{-ax^2}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}} \quad (\mathrm{Re}(a) > 0), \qquad I_M = \int_{-\infty}^{\infty} e^{\mathrm{i}ax^2}\,\mathrm{d}x.$$

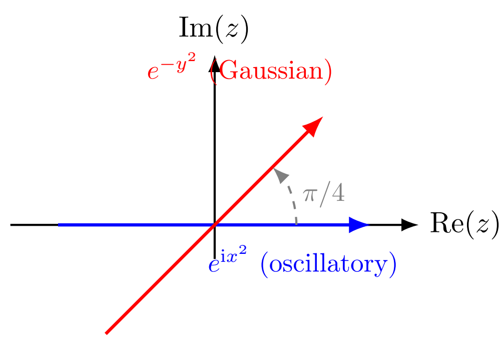
The Euclidean integral I_E converges *absolutely*: the integrand $|e^{-ax^2}| = e^{-\mathrm{Re}(a)\,x^2}$ decays as $x \to \pm\infty$, and the integral is dominated by $|x| \lesssim 1/\sqrt{a}$. It defines a probability measure: $p(x) \propto e^{-ax^2} \geq 0$.

The Minkowski integral I_M is *not* absolutely convergent: $|e^{\mathrm{i}ax^2}| = 1$ for all x. The integral converges only by cancellation of phases — the rapid oscillation of $e^{\mathrm{i}ax^2}$ at large $|x|$ causes positive and negative contributions to cancel. It does not define a probability measure.

To evaluate I_M, we *rotate the contour* in the complex z-plane. Set $z = e^{\mathrm{i}\pi/4}y$ so that $z^2 = \mathrm{i}y^2$:

$$\int_{-\infty}^{\infty} e^{\mathrm{i}z^2}\,\mathrm{d}z = e^{\mathrm{i}\pi/4}\int_{-\infty}^{\infty} e^{-y^2}\,\mathrm{d}y = e^{\mathrm{i}\pi/4}\sqrt{\pi}.$$

The contour rotation transformed the oscillatory integrand $e^{\mathrm{i}x^2}$ into the Gaussian integrand e^{-y^2}. This is Wick rotation in its simplest form: a 90° contour rotation that converts oscillation into decay.



44.2 The Lorentzian Path Integral

In real-time quantum mechanics, transition amplitudes are computed by the **Lorentzian** (or Minkowski) path integral:

$$\langle \phi_f | e^{-\mathrm{i}Ht} | \phi_i \rangle = \int_{\phi(0)=\phi_i}^{\phi(t)=\phi_f} \mathcal{D}\phi \ \exp(\mathrm{i} S_M[\phi]) \,, \tag{44.1}$$

where $S_M[\phi] = \int \mathrm{d}t \ \mathrm{d}^d x \, \mathcal{L}(\phi, \partial_t \phi, \nabla \phi)$ is the **Minkowski action** and $\mathcal{L}$ is the Lagrangian density. For a scalar field, a typical Lagrangian is

$$\mathcal{L} = \frac{1}{2}(\partial_t \phi)^2 - \frac{1}{2}(\nabla \phi)^2 - V(\phi).$$

The crucial feature is the factor of i in $e^{\mathrm{i}S_M}$: the weight is a *phase*, not a positive real number. The integrand oscillates rather than decays, and the path integral is defined by destructive interference among paths far from the classical solution, rather than by exponential suppression. This is fundamentally different from the Euclidean path integral, where e^{-S_E} is positive and defines a probability measure.

44.3 Wick Rotation

The connection between the Lorentzian and Euclidean formulations is the **Wick rotation**: the substitution $t = -\mathrm{i}\tau$, where τ is the Euclidean (imaginary) time. Under this substitution:

$$\mathrm{i} \ \mathrm{d}t = \mathrm{d}\tau, \qquad (\partial_t \phi)^2 \to -(\partial_\tau \phi)^2, \qquad \mathrm{i}S_M \to -S_E,$$

so $e^{\mathrm{i}S_M} \to e^{-S_E}$. The oscillatory Lorentzian weight becomes a real, positive Euclidean weight.

Let us see this explicitly. The Minkowski action for a free scalar field is

$$S_M = \int \mathrm{d}t \ \mathrm{d}^d x \left[\frac{1}{2}(\partial_t \phi)^2 - \frac{1}{2}(\nabla \phi)^2 - \frac{1}{2} m^2 \phi^2 \right].$$

Setting $t = -\mathrm{i}\tau$ so that $\partial_t = \mathrm{i}\partial_\tau$ and $\mathrm{d}t = -\mathrm{i}\,\mathrm{d}\tau$:

$$\mathrm{i}S_M = \mathrm{i}(-\mathrm{i}) \int \mathrm{d}\tau \ \mathrm{d}^d x \left[\frac{1}{2}(-1)(\partial_\tau \phi)^2 - \frac{1}{2}(\nabla \phi)^2 - \frac{1}{2} m^2 \phi^2 \right] = -\int \mathrm{d}\tau \ \mathrm{d}^d x \left[\frac{1}{2}(\partial_\tau \phi)^2 + \frac{1}{2}(\nabla \phi)^2 + \frac{1}{2} m^2 \phi^2 \right].$$

The right-hand side is $-S_E$, where

$$S_E = \int \mathrm{d}\tau \ \mathrm{d}^d x \left[\frac{1}{2}(\partial_\tau \phi)^2 + \frac{1}{2}(\nabla \phi)^2 + \frac{1}{2} m^2 \phi^2 \right]. \tag{44.2}$$

Notice that all signs are positive in S_E: the Euclidean action is a sum of squares, which makes e^{-S_E} a well-defined (positive, normalizable) weight. The Lorentzian action has a relative minus sign between the time and space derivatives (reflecting the $(-,+,+,+)$ signature of Minkowski spacetime), which makes $e^{\mathrm{i}S_M}$ oscillatory.

The Wick rotation is not merely a formal substitution. Under appropriate analyticity conditions (related to **reflection positivity** in the Euclidean formulation and the **spectrum condition** in the Lorentzian formulation), the two path integrals contain equivalent physical information. The Euclidean formulation is better suited for computation (the integral converges), while the Lorentzian formulation is closer to the physics of real-time evolution and measurement.

The precise conditions under which this equivalence holds, and the complex-analytic machinery that underlies it, are the subject of the next chapter.

44.4 The Transfer Matrix: Why Wick Rotation Is a Theorem

On a lattice, the equivalence between Lorentzian and Euclidean formulations is not an assumption—it is a theorem. The foundational object is the **transfer matrix** $\hat{T} = e^{-Ha}$, where H is the Hamiltonian and a is the lattice spacing in the time direction. The Euclidean path integral on a lattice with N_τ time slices is

$$Z_E = \mathrm{Tr}[\hat{T}^{N_\tau}] = \sum_k e^{-E_k N_\tau a},$$

where E_k are the eigenvalues of H. The path integral is a *derived identity*: inserting a complete set of field eigenstates between each pair of adjacent transfer matrices produces $Z_E = \int \prod_{n=1}^{N_\tau} \mathrm{d}\phi_n \, e^{-S_E[\phi]}$, where the lattice action S_E emerges from the matrix elements of $\hat{T}$. The "functional measure" $\mathcal{D}\phi = \prod_n \mathrm{d}\phi_n$ is a finite product of ordinary Lebesgue measures—well-defined without any appeal to functional analysis on infinite-dimensional spaces.

The Lorentzian evolution operator $e^{-\mathrm{i}Ht}$ and the Euclidean one $e^{-H\tau}$ are the *same operator* evaluated at different values of a complex parameter. The analyticity that makes the Wick rotation work is a consequence of **spectral positivity**: because $E_k \geq 0$ (the Hamiltonian is bounded below), the function $\mathrm{Tr}[e^{-Hz}]$ is analytic in the right half-plane $\mathrm{Re}(z) > 0$.

Property	**Lorentzian**	**Euclidean**
Time parameter	Real t	Imaginary $\tau = \mathrm{i}t$
Evolution operator	Unitary: $e^{-\mathrm{i}Ht}$	Contraction: $e^{-H\tau}$
Path integral weight	$e^{\mathrm{i}S_M}$ (complex, $\lvert e^{\mathrm{i}S}\rvert = 1$)	e^{-S_E} (real, positive)
Correlator	Oscillatory spectral sum	Exponential decay sum
Mass extraction	Pole: $1/(p^2 - m^2 + \mathrm{i}\epsilon)$	Decay rate: $G_E(\tau) \sim e^{-m\tau}$
Probability measure	No	Yes (e^{-S_E}/Z)
Monte Carlo	Impossible	Standard (MCMC)
Convergence	Conditional (by cancellation)	Absolute

44.5 The Quantum–Classical Correspondence

A remarkable consequence of the Euclidean path integral: the partition function of a *quantum* system in d spatial dimensions at finite temperature is equal to the partition function of a *classical* statistical system in $d{+}1$ dimensions, where the extra dimension (Euclidean time) is compact with circumference $\beta = 1/T$.

Quantum (d dimensions)	**Classical ($d{+}1$ dimensions)**
Temperature $T = 1/\beta$	Lattice extent $L_\tau = \beta$
Zero temperature ($T \to 0$)	Infinite extent ($L_\tau \to \infty$)
Quantum ground state	Classical ground state
Energy gap ΔE	Correlation length $\xi = 1/\Delta E$
Quantum phase transition ($T = 0$)	Classical critical point

This correspondence is not a metaphor; it is an exact mathematical identity. It is the reason why methods from classical statistical mechanics (Monte Carlo, renormalization group, conformal field theory) apply to quantum field theory, and vice versa.

44.6 Three Roles of Euclidean Time

Euclidean time τ plays three different roles depending on context:

Analytic continuation parameter. In the vacuum theory ($T = 0$, $\tau \in (-\infty, \infty)$), Euclidean time is a mathematical device for making the path integral convergent. The physical content is extracted by analytic continuation back to real time.

Inverse temperature. In the thermal theory ($T > 0$, $\tau \in [0, \beta]$ with periodic boundary conditions for bosons, antiperiodic for fermions), Euclidean time *is* the physical inverse temperature. The Euclidean path integral with periodic boundary conditions *is* the thermal partition function $Z(\beta) = \mathrm{Tr}(e^{-\beta H})$. No analytic continuation is needed.

RG "time." In the Wilsonian RG, the logarithm of the momentum scale plays the role of a "time" variable, and the RG flow equations are "evolution equations" in this time. This is purely Euclidean; there is no Lorentzian analogue.

The Euclidean formulation is a *lossy compression* of the Lorentzian theory: it preserves the spectrum (masses, energy gaps) but discards phase information (interference, real-time evolution, superposition). The information that is lost can in principle be recovered by analytic continuation, but this recovery is exponentially ill-conditioned in practice.

Chapter 45

Contour Methods, Poles, and the iϵ Prescription

45.1 Why Complex Analysis Enters Physics

The Lorentzian path integral has the weight $e^{\mathrm{i}S_M}$, which is oscillatory. To evaluate integrals involving oscillatory integrands, we exploit a fundamental principle: an oscillatory integral over the real line can often be computed by deforming the integration contour into the complex plane, where the integrand decays exponentially and the integral converges manifestly.

This is the reason complex analysis pervades quantum field theory. The physical quantities are defined on the real axis (real frequencies, real times), but the *computation* of those quantities often requires a detour into the complex plane. The Wick rotation is the most dramatic example: we rotate the time contour from the real axis to the imaginary axis, converting the oscillatory Lorentzian weight into the decaying Euclidean weight.

45.2 Cauchy's Theorem and Contour Integration

The central result of complex analysis that makes all of this work is Cauchy's theorem. Let $f(z)$ be a function that is analytic (holomorphic) inside and on a closed contour $\mathcal{C}$ in the complex plane. Then

$$\oint_{\mathcal{C}} f(z)\,\mathrm{d}z = 0. \tag{45.1}$$

If f has isolated poles inside $\mathcal{C}$, the integral picks up contributions from those poles via the **residue theorem**:

$$\oint_{\mathcal{C}} f(z)\,\mathrm{d}z = 2\pi\mathrm{i} \sum_{\text{poles inside } \mathcal{C}} \mathrm{Res}(f, z_k). \tag{45.2}$$

The crucial consequence is that if $f(z)$ is analytic in a region, the integral of f along any contour within that region depends only on the endpoints (and on any poles enclosed). This means we can *deform* the contour without changing the value of the integral, as long as we do not cross any poles or branch cuts.

45.3 Poles of the Propagator and the iϵ Prescription

Consider the free retarded propagator for a scalar field, which we need to evaluate the integral

$$G(t, \mathbf{x}) = \int \frac{\mathrm{d}\omega}{2\pi} \frac{e^{-\mathrm{i}\omega t}}{\omega^2 - \omega_{\mathbf{k}}^2}.$$

The integrand has poles at $\omega = \pm\omega_{\mathbf{k}}$ on the real axis. The integral is ambiguous as written: the contour passes directly through the poles. To define it, we must specify how the contour avoids the poles—and different prescriptions correspond to different physical boundary conditions.

The iϵ **prescription** displaces the poles slightly off the real axis:

Retarded propagator (G_R): both poles are pushed below the real axis, $\omega \to \omega + \mathrm{i}\epsilon$:

$$G_R(\omega, \mathbf{k}) = \frac{1}{(\omega + \mathrm{i}\epsilon)^2 - \omega_{\mathbf{k}}^2}. \tag{45.3}$$

For $t > 0$, we close the contour in the lower half-plane (where $e^{-\mathrm{i}\omega t}$ decays), enclosing both poles and getting a nonzero result. For $t < 0$, we close in the upper half-plane, enclosing no poles, giving $G_R(t < 0) = 0$. The retarded propagator vanishes for $t < 0$: cause precedes effect. This is the boundary condition appropriate for response functions.

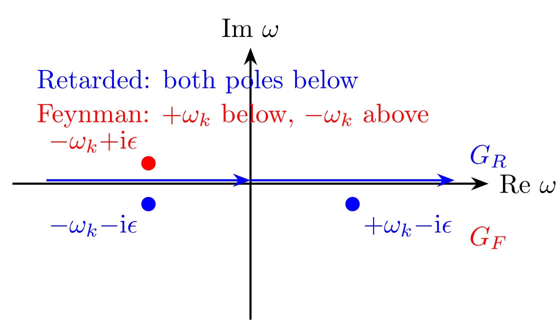
Advanced propagator (G_A): both poles are pushed above the real axis, $\omega \to \omega - \mathrm{i}\epsilon$. This gives $G_A(t > 0) = 0$—the time-reversed boundary condition.

Feynman propagator (G_F): the positive-energy pole is pushed below and the negative-energy pole above:

$$G_F(\omega, \mathbf{k}) = \frac{1}{\omega^2 - \omega_{\mathbf{k}}^2 + \mathrm{i}\epsilon}. \tag{45.4}$$

This prescription selects the vacuum boundary condition—positive-frequency modes propagate forward in time, negative-frequency modes propagate backward. It is the natural propagator for scattering problems and for the path-integral formulation of quantum field theory.

The three prescriptions give the same propagator in the Euclidean formulation (after Wick rotation), because the Euclidean propagator $1/(\omega_n^2 + \omega_{\mathbf{k}}^2)$ has no poles on the imaginary axis. The distinction between retarded, advanced, and Feynman arises only in the Lorentzian formulation, where the poles sit on or near the real axis.



45.4 The Contour Rotation That Gives Wick Rotation

The Wick rotation can now be understood as a contour deformation in the complex ω-plane (or equivalently, the complex t-plane).

Consider a frequency integral along the real ω-axis. If the integrand (including the propagator and the $e^{-i\omega t}$ factor) is analytic in the first and third quadrants of the complex ω-plane—that is, if it has no poles or branch cuts in those quadrants—then we can rotate the contour by 90°:

$$\omega_{\text{real}} \to \omega_E = i\omega_{\text{real}},$$

mapping the real-frequency integral to an integral along the imaginary axis. Setting $\omega = i\omega_E$ with ω_E real:

$$\omega^2 = -\omega_E^2, \qquad \omega^2 - \omega_{\mathbf{k}}^2 \to -\omega_E^2 - \omega_{\mathbf{k}}^2,$$

and the Lorentzian propagator $1/(\omega^2 - \omega_{\mathbf{k}}^2)$ becomes the Euclidean propagator $-1/(\omega_E^2 + \omega_{\mathbf{k}}^2)$ (up to sign). The oscillatory factor $e^{-i\omega t}$ becomes $e^{-\omega_E \tau}$ (with τ the Euclidean time), which decays exponentially—the integral now converges manifestly.

The condition for this rotation to be valid is precisely the condition that no poles are crossed during the rotation. For the retarded propagator, both poles are in the lower half-plane, and the rotation from real axis to positive imaginary axis passes through the upper half-plane, which is pole-free. For the Feynman propagator, the pole at $-\omega_{\mathbf{k}} + i\epsilon$ is in the upper half-plane, so one must be more careful—but the $i\epsilon$ displacement ensures the pole is avoided.

45.5 Analyticity and Its Physical Meaning

The analyticity properties of propagators and response functions are not mathematical accidents. They encode fundamental physical principles:

Causality requires that the retarded propagator $G_R(\omega)$ is analytic in the upper half of the complex ω-plane. This is because $G_R(t) = 0$ for $t < 0$ (no effect before cause), and a function that vanishes for $t < 0$ has a Fourier transform that is analytic in the upper half-plane (by the Paley–Wiener theorem).

The spectrum condition (positivity of energy) requires that the spectral function $\rho(\omega) \geq 0$. Combined with the spectral representation (47.1), this determines the analytic structure: $G_R(\omega)$ has all its singularities (poles and branch cuts) on or below the real axis, and approaches $1/\omega^2$ at large $|\omega|$.

Stability requires that the poles of the retarded propagator are in the lower half-plane ($\mathrm{Im}\,\omega < 0$), corresponding to modes that decay in time rather than grow. A pole in the upper half-plane would signal an instability.

These properties—causality, spectral positivity, stability—are the physical guarantees that make the Wick rotation trustworthy.

45.6 When Is the Wick Rotation Trusted?

The Wick rotation is reliable when the following conditions hold:

No poles in the rotation quadrant. The integrand must be analytic in the region swept by the contour rotation. For the retarded propagator at zero density, this is guaranteed by causality.

Decay at infinity. The integrand must vanish fast enough on the quarter-circle at infinity to contribute nothing. This is typically ensured by the convergence of the Euclidean integral.

Reflection positivity (or, equivalently, the Osterwalder–Schrader axioms) in the Euclidean formulation guarantees that the Euclidean theory can be reconstructed into a unitary quantum theory.

The Wick rotation can *fail* or require modification in several important cases:

Finite-density systems. A nonzero chemical potential μ shifts the poles of the fermion propagator: $G^{-1} = \mathrm{i}\omega_n - (\xi_{\mathbf{k}} - \mu)$. In Minkowski space, the Fermi surface introduces poles near the real axis that obstruct the naive contour rotation.

Real-time dynamics. Computing genuinely out-of-equilibrium quantities (transport coefficients, relaxation rates, real-time response) requires the Lorentzian formulation directly, since the Euclidean formulation only accesses equilibrium information. Analytic continuation from Euclidean data is formally exact but numerically ill-conditioned.

Systems with a sign problem. When the Euclidean action is complex (finite baryon density in QCD, systems with a θ-term), the weight e^{-S_E} is not positive, and the probabilistic interpretation fails. The Wick rotation still defines a formal integral, but it does not define a probability measure.

Gravity and cosmology. In gravitational path integrals, the conformal factor problem means that the Euclidean action is unbounded below, and the Wick rotation requires a complex contour (the Gibbons–Hawking–Perry contour) rather than a simple 90° rotation.

In the contexts treated in this book—non-relativistic condensed matter at finite temperature, and relativistic field theory in equilibrium—the Wick rotation is on solid ground, and the Euclidean formulation is the natural computational framework.

Chapter 46

The Euclidean–Minkowski Correspondence for Propagators

46.1 The Free Euclidean Propagator

In the Euclidean formulation, the free propagator in momentum-frequency space is

$$G_E(\omega_n, \mathbf{k}) = \frac{1}{\omega_n^2 + \mathbf{k}^2 + m^2}, \tag{46.1}$$

where ω_n are Matsubara frequencies ($\omega_n = 2n\pi T$ for bosons, $(2n{+}1)\pi T$ for fermions at temperature T). In Euclidean position space, the propagator decays exponentially:

$$G_E(\tau, \mathbf{x}) \sim e^{-m\sqrt{\tau^2+\mathbf{x}^2}} \quad \text{at large separations.}$$

The mass m is the inverse of the correlation length: it controls the rate of exponential decay.

46.2 Analytic Continuation to the Retarded Propagator

The **retarded propagator**—the real-time response function—is obtained by analytic continuation:

$$G_E(\mathrm{i}\omega_n, \mathbf{k})\Big|_{\mathrm{i}\omega_n \to \omega + \mathrm{i}0^+} = G_R(\omega, \mathbf{k}). \tag{46.2}$$

For the free field, substituting $\mathrm{i}\omega_n \to \omega + \mathrm{i}0^+$ in (46.1):

$$G_R(\omega, \mathbf{k}) = \frac{1}{-(\omega + \mathrm{i}0^+)^2 + \mathbf{k}^2 + m^2} = \frac{-1}{(\omega + \mathrm{i}0^+)^2 - \omega_{\mathbf{k}}^2},$$

where $\omega_{\mathbf{k}} = \sqrt{\mathbf{k}^2 + m^2}$ is the **dispersion relation**. This has poles at $\omega = \pm\omega_{\mathbf{k}} - \mathrm{i}0^+$—just below the real axis, as required for a retarded function.

46.3 The Correspondence: Decay Rate vs. Pole

The key correspondence between the Euclidean and Minkowski formulations is:

Euclidean quantity	**Minkowski quantity**
Exponential decay rate in τ: $G_E \sim e^{-m\tau}$	Energy of excitation: $\omega_{\mathbf{k}} = \sqrt{\mathbf{k}^2 + m^2}$
Correlation length $\xi = 1/m$	Compton wavelength $\lambda_C = 1/m$
Gap in Euclidean spectrum	Mass of the particle
Decay rate of Euclidean correlator	Pole of retarded propagator

In the Euclidean formulation, the mass m appears as a decay rate: the two-point function $G_E(\tau) \sim e^{-m|\tau|}$ falls off exponentially in imaginary time, and the rate of falloff is the mass. In the Minkowski formulation, the same m appears as the location of a pole: the retarded propagator $G_R(\omega)$ has a pole at $\omega = \omega_{\mathbf{k}} = \sqrt{\mathbf{k}^2 + m^2}$, and this pole determines the energy of the excitation.

This is the deep reason why the "gap" we computed throughout this book—a parameter controlling the exponential decay of Euclidean correlators—has the physical interpretation of a particle mass or excitation energy. The Wick rotation maps one into the other.

Chapter 47

Quasiparticle Certification: The Width Γ and Its Meaning

47.1 The Spectral Function

For an interacting system, the retarded propagator has the **spectral representation**

$$G_R(\omega, \mathbf{k}) = \int_{-\infty}^{\infty} \frac{\rho(\omega', \mathbf{k})}{\omega - \omega' + \mathrm{i}0^+} \, \mathrm{d}\omega', \tag{47.1}$$

where $\rho(\omega, \mathbf{k}) \geq 0$ is the **spectral function** (for bosonic fields; the positivity follows from reflection positivity or, equivalently, from unitarity). The spectral function satisfies the sum rule $\int \rho(\omega, \mathbf{k}) \, \mathrm{d}\omega = 1$ (with appropriate normalization).

The spectral function can also be expressed in terms of the self-energy. If the retarded propagator has the form $G_R(\omega, \mathbf{k}) = 1/(\omega^2 - \omega_{\mathbf{k}}^2 - \Sigma_R(\omega, \mathbf{k}))$ (for a relativistic scalar, with suitable conventions), then

$$\rho(\omega, \mathbf{k}) = -\frac{1}{\pi} \frac{\operatorname{Im} \Sigma_R(\omega, \mathbf{k})}{(\omega^2 - \omega_{\mathbf{k}}^2 - \operatorname{Re} \Sigma_R)^2 + (\operatorname{Im} \Sigma_R)^2}. \tag{47.2}$$

This is the general form. The shape of the spectral function is determined by the real and imaginary parts of the self-energy.

47.2 The Quasiparticle Pole

If the self-energy is small (weak interactions), the spectral function is sharply peaked. Near the peak at $\omega \approx \Omega_{\mathbf{k}}$ (the renormalized energy), we can expand:

$$\omega^2 - \omega_{\mathbf{k}}^2 - \Sigma_R(\omega, \mathbf{k}) \approx 2\Omega_{\mathbf{k}}(\omega - \Omega_{\mathbf{k}}) + \mathrm{i}\,\Gamma_{\mathbf{k}},$$

where $\Omega_{\mathbf{k}}$ solves $\Omega_{\mathbf{k}}^2 = \omega_{\mathbf{k}}^2 + \operatorname{Re} \Sigma_R(\Omega_{\mathbf{k}}, \mathbf{k})$ (the renormalized dispersion) and $\Gamma_{\mathbf{k}} = -\operatorname{Im} \Sigma_R(\Omega_{\mathbf{k}}, \mathbf{k})/(2\Omega_{\mathbf{k}})$ is the **damping rate**. The spectral function near the peak becomes a Lorentzian:

$$\rho(\omega, \mathbf{k}) \approx \frac{Z_{\mathbf{k}}}{\pi} \frac{\Gamma_{\mathbf{k}}}{(\omega - \Omega_{\mathbf{k}})^2 + \Gamma_{\mathbf{k}}^2}, \tag{47.3}$$

where $Z_{\mathbf{k}} \leq 1$ is the **quasiparticle residue** (the fraction of the total spectral weight carried by the peak). The remainder $1 - Z_{\mathbf{k}}$ is distributed in a broad, incoherent background.

47.3 Deriving Γ from the Self-Energy

The damping rate $\Gamma_{\mathbf{k}}$ has a direct physical derivation. The imaginary part of the retarded self-energy, $\operatorname{Im}\Sigma_R(\omega, \mathbf{k})$, arises from on-shell processes—scattering events that conserve energy and momentum. By the optical theorem (a consequence of unitarity), $\operatorname{Im}\Sigma_R$ is proportional to the total rate of all processes that can scatter the quasiparticle out of its initial state:

$$\Gamma_{\mathbf{k}} = -\frac{\operatorname{Im}\Sigma_R(\Omega_{\mathbf{k}}, \mathbf{k})}{2\Omega_{\mathbf{k}}}.$$

In a Fermi liquid at low temperature, the leading contribution to $\operatorname{Im}\Sigma_R$ comes from two-particle scattering, and phase-space restrictions give $\Gamma_{\mathbf{k}} \propto (\omega - \mu)^2 + (\pi T)^2$. This quadratic vanishing at the Fermi surface is the reason why Landau's Fermi liquid theory works: near the Fermi surface, quasiparticles are long-lived because the phase space for scattering vanishes.

47.4 The Quasiparticle Criterion

A quasiparticle is well-defined when the peak in the spectral function is sharp—that is, when the width $\Gamma_{\mathbf{k}}$ is much smaller than the energy $\Omega_{\mathbf{k}}$:

$$\boxed{\Gamma_{\mathbf{k}} \ll \Omega_{\mathbf{k}} \qquad \text{(well-defined quasiparticle).}} \tag{47.4}$$

When this condition holds, the excitation oscillates many times before decaying, and it makes sense to assign it a definite energy $\Omega_{\mathbf{k}}$ and to think of it as a "particle" propagating through the medium with a well-defined dispersion relation.

When $\Gamma_{\mathbf{k}} \sim \Omega_{\mathbf{k}}$, the peak is broad and merges with the incoherent background. The excitation decays before completing even one oscillation, and the quasiparticle description breaks down. The spectral function is no longer a sharp peak but a broad bump, and the concept of a particle with a definite energy loses its meaning.

In the Euclidean formulation, the same criterion has a clear signature: a well-defined quasiparticle produces an exponentially decaying correlator $G_E(\tau) \sim Z\, e^{-\Omega|\tau|}$ at intermediate times, while a broad spectral function produces power-law or non-exponential decay.

Chapter 48

Quasiparticle Modes vs. Hydrodynamic Modes

48.1 Two Kinds of Long-Lived Excitations

Not all long-lived excitations are quasiparticles, and not all sharp features in response functions arise from the quasiparticle mechanism. There are two fundamentally different reasons why a mode can be long-lived:

Quasiparticle modes are long-lived because the *damping rate* Γ *is small compared to the frequency* Ω. This happens when the self-energy has a small imaginary part, typically because phase space for decay is restricted (as in a Fermi liquid near the Fermi surface, or a gapped system below the gap). The mode exists at all wavelengths (including short wavelengths), and its lifetime is controlled by microscopic scattering processes.

Hydrodynamic modes are long-lived because they exist *at long wavelengths, where both* Ω *and* Γ *vanish together.* Their existence is guaranteed by conservation laws (Chapter 23), not by weak scattering. At any nonzero wavevector, a hydrodynamic mode has a definite $\Gamma(q)$ that does not vanish—but the ratio Γ/Ω can remain small. The mode becomes exact only at $q = 0$, where it corresponds to a conserved quantity.

The distinction is best understood through an example.

48.2 Phonons: From Quasiparticle to Hydrodynamic

Consider sound waves (phonons) in a crystal or a fluid. The dispersion relation is $\Omega(q) = c\,q$ at long wavelengths, where c is the speed of sound. The damping rate, due to anharmonic scattering and viscous processes, scales as $\Gamma(q) \propto q^2$ (in three dimensions, at low temperature).

At long wavelengths, the ratio $\Gamma/\Omega \propto q^2/(cq) = q/c \to 0$ as $q \to 0$. Sound waves become increasingly well-defined at long wavelengths. This is the hydrodynamic regime: the mode is long-lived because conservation of momentum and energy constrains the system to oscillate rather than relax.

At short wavelengths (comparable to the mean free path ℓ), the ratio Γ/Ω grows, and eventually $\Gamma \sim \Omega$. Here the phonon ceases to be a well-defined excitation. The spectral function broadens, and the quasiparticle description breaks down.

In between, at intermediate wavelengths where $q\ell \lesssim 1$ but $\Gamma \ll \Omega$, the phonon is a well-defined quasiparticle in the sense of (47.4). It propagates ballistically, scattering occasionally off other phonons or defects, with a lifetime $\tau_{\mathbf{k}} = 1/(2\Gamma_{\mathbf{k}})$.

This illustrates the crossover:

Regime	**Condition**	**Character**
Hydrodynamic	$q\ell \ll 1$	Sound: $\omega = cq - \mathrm{i}\Gamma_{\text{visc}} q^2$
Quasiparticle	$q\ell \sim 1$, $\Gamma \ll \Omega$	Ballistic phonon with finite lifetime
Overdamped	$\Gamma \sim \Omega$	Broad spectral feature, no particle

48.3 The Mathematical Distinction

Mathematically, the two kinds of modes correspond to different structures in the inverse propagator (precision):

For a **quasiparticle mode**, the inverse propagator has a pole structure

$$G^{-1}(\omega, \mathbf{k}) = \omega^2 - \omega_{\mathbf{k}}^2 - \Sigma(\omega, \mathbf{k}),$$

where $\omega_{\mathbf{k}}^2 = \mathbf{k}^2 + m^2$ is determined by the dispersion and Σ is small. The mode exists because the *bare* dispersion already has a pole, and interactions only broaden it.

For a **hydrodynamic mode**, the inverse propagator has the form

$$K(\omega, q) = -\mathrm{i}\omega\chi + D_{\text{diff}}\chi q^2 + \cdots \quad \text{(diffusion)},$$

or

$$K(\omega, q) = \chi\omega^2 - \rho_s q^2 + \mathrm{i}\eta\omega q^2 + \cdots \quad \text{(sound)},$$

where the structure is dictated by conservation laws and symmetry (Chapter 23). The mode exists because conservation *forces* the inverse propagator to vanish at $\omega = q = 0$, regardless of the strength of interactions. The damping terms ($\mathrm{i}\eta\omega q^2$) arise from dissipative processes and are determined by transport coefficients (viscosity, thermal conductivity), not by microscopic scattering rates.

48.4 Why the Distinction Matters

The distinction between quasiparticle and hydrodynamic modes has practical consequences. Quasiparticle modes can cease to exist if interactions become too strong (the spectral function broadens until $\Gamma \sim \Omega$). Hydrodynamic modes cannot cease to exist as long as the conservation law holds: they are protected by symmetry. Even in a strongly interacting system where no quasiparticles survive, sound and diffusion persist.

This is why hydrodynamics works even when the quasiparticle description fails. A quark-gluon plasma at high temperature has no well-defined quasiparticles (the coupling is strong), but it still has sound waves, diffusion of conserved charges, and viscous flow—all hydrodynamic modes protected by conservation of energy-momentum and conserved charges.

What was exact, what was approximate, and what was conventional?

Exact. The spectral representation (47.1) and positivity of ρ are exact consequences of unitarity (or reflection positivity). The existence of hydrodynamic modes from conservation laws is exact and model-independent.

Conditional. The Wick rotation is an exact equivalence, but only under stated conditions: reflection positivity in the Euclidean theory, spectral positivity (bounded-below Hamiltonian) in the Lorentzian theory, and no poles in the rotation quadrant. These conditions can fail at finite density (complex action), in gravity (conformal-factor problem), and in systems with a sign problem. The analytic continuation from discrete Matsubara data to real frequencies is formally unique (by the Carlson theorem, given suitable growth bounds), but it is an exponentially ill-conditioned inverse problem: small errors in the Euclidean data produce large errors in the reconstructed spectral function. This is why numerical analytic continuation (e.g., via the maximum entropy method or Nevanlinna analytic continuation) requires prior assumptions about the spectral function's smoothness.

Approximate. The Lorentzian form (47.3) for the spectral function assumes weak damping ($\Gamma \ll \Omega$). The expression for Γ in terms of $\operatorname{Im}\Sigma_R$ assumes the self-energy varies slowly near the pole. The scaling $\Gamma \propto q^2$ for phonon damping is a leading-order result in the gradient expansion. Analytic continuation from discrete Matsubara frequencies to real frequencies is formally exact but numerically ill-conditioned—in practice, it requires additional assumptions about the spectral function.

Conventional. The terms "quasiparticle," "spectral function," "damping rate," "quasiparticle residue," and "hydrodynamic mode" are standard physics vocabulary.

48.5 The Kramers–Kronig Relations

The analyticity of the retarded propagator $G_R(\omega)$ in the upper half-plane, which follows from causality ($G_R(t<0)=0$), implies powerful sum rules relating the real and imaginary parts of G_R.

Starting from the spectral representation $G_R(\omega) = \int \frac{\rho(\omega')}{\omega-\omega'+\mathrm{i}0^+}\,\mathrm{d}\omega'$, separate real and imaginary parts using $\frac{1}{x+\mathrm{i}0^+} = \mathcal{P}\frac{1}{x} - \mathrm{i}\pi\delta(x)$:

$$\operatorname{Re} G_R(\omega) = \mathcal{P}\int \frac{\rho(\omega')}{\omega-\omega'}\,\mathrm{d}\omega', \qquad \operatorname{Im} G_R(\omega) = -\pi\rho(\omega).$$

Substituting the second into the first:

$$\boxed{\operatorname{Re} G_R(\omega) = -\frac{1}{\pi}\mathcal{P}\int_{-\infty}^{\infty} \frac{\operatorname{Im} G_R(\omega')}{\omega-\omega'}\,\mathrm{d}\omega'.} \tag{48.1}$$

This is the first **Kramers–Kronig relation**: the real part of any causal response function is determined by its imaginary part (and vice versa, by exchanging real and imaginary parts).

The Kramers–Kronig relations are model-independent—they follow solely from causality and the assumption that $G_R(\omega) \to 0$ as $|\omega| \to \infty$ (fast enough for the dispersion integral to converge). They provide exact sum rules that any approximation must satisfy. For example, setting $\omega = 0$ in (48.1):

$$\operatorname{Re} G_R(0) = \frac{1}{\pi}\int_0^{\infty} \frac{2\omega' \operatorname{Im} G_R(\omega')}{\omega'^2}\,\mathrm{d}\omega',$$

which relates the static (zero-frequency) response to the integrated spectral weight—a useful check on any calculation.

For the dielectric function $\varepsilon(\omega) = 1 + V(q)\Pi_R(\omega,q)$, the Kramers–Kronig relations give

$$\operatorname{Re}\varepsilon(\omega) - 1 = \frac{2}{\pi}\mathcal{P}\int_0^{\infty} \frac{\omega' \operatorname{Im}\varepsilon(\omega')}{\omega'^2-\omega^2}\,\mathrm{d}\omega',$$

which is the standard form used in optics and condensed-matter physics. The f-sum rule $\int_0^\infty \omega \operatorname{Im}\varepsilon(\omega)\,\mathrm{d}\omega = \pi\omega_p^2/2$ constrains the total oscillator strength and is an important consistency check on calculations of ε.

Chapter 49

Unitary Evolution and the Born Rule

49.1 What the Euclidean Formulation Does and Does Not Compute

The Euclidean path integral gives us: *thermodynamic quantities* (free energy $F = -T \log Z$, internal energy, entropy); *imaginary-time correlation functions* $G_E(\tau, \mathbf{x})$, from which we can extract energy gaps and spatial correlation lengths; *ground-state properties* (in the $\beta \to \infty$ limit); and *spectral information* through analytic continuation.

It does not directly give: *real-time dynamics* (time evolution of observables, relaxation of non-equilibrium states); *individual measurement probabilities* (the Born rule involves real-time amplitudes, not Euclidean correlators); or *interference and superposition* (which arise from the phase of complex amplitudes, not visible in the Euclidean path integral where all weights are real and positive).

The probability of observing outcome a is given by the **Born rule**: $P(a) = |\langle a|\psi(t)\rangle|^2$. The quantum state encodes amplitudes—complex numbers whose squared moduli give probabilities. The amplitudes evolve unitarily; the probabilities are derived from them through the Born rule.

49.2 The Role of Physical Stories

Physical stories about particles, quasiparticles, virtual processes, and collective modes are interpretations of spectral and correlation structures. They are vivid, useful, and sometimes essential for guiding intuition and designing experiments. But they should be recognized as stories—interpretive frameworks that illuminate the mathematics, not replacements for it.

The Euclidean formulation is a probability theory. The real-time formulation is a theory of amplitudes. The Born rule bridges the two. This book has developed the Euclidean side using the language of probability and statistics. The reader who wishes to go further will need to engage with the amplitude structure directly. But the Euclidean framework provides a solid mathematical foundation from which to approach these deeper questions.

49.3 Limits of the Probability Interpretation

We close by noting explicitly where the probability interpretation breaks down: Grassmann variables are not ordinary random variables; fermion determinants may fail to be positive (the fermion sign problem); finite-density systems may have a sign problem; gauge theories require constraints or gauge fixing that modify the integration measure; complex actions do not define ordinary probability measures. We have used probability language where it is mathematically legitimate, and flagged explicitly when it is being used as an analogy or formal extension.

49.4 Concrete Example: The Two-Level System

To make the Euclidean–Lorentzian correspondence tangible, consider the simplest possible quantum system: a particle that can be in one of two states, $|0\rangle$ (energy 0) and $|1\rangle$ (energy E). The Hamiltonian is $H = E|1\rangle\langle 1|$, and the transfer matrix is $\hat{T} = e^{-Ha} = |0\rangle\langle 0| + e^{-Ea}|1\rangle\langle 1|$.

The Euclidean two-point function is

$$G_E(\tau) = \frac{\mathrm{Tr}(e^{-(\beta-\tau)H}\hat{O}\, e^{-\tau H}\hat{O})}{\mathrm{Tr}\; e^{-\beta H}},$$

where $\hat{O} = |0\rangle\langle 1| + |1\rangle\langle 0|$ is the transition operator. At zero temperature ($\beta \to \infty$), the ground state $|0\rangle$ dominates and $G_E(\tau) = e^{-E\tau}$: the Euclidean correlator decays exponentially with rate E. The "mass" (in the language of this book) is $m = E$.

The retarded propagator, obtained by analytic continuation, is

$$G_R(\omega) = \frac{1}{\omega - E + \mathrm{i}0^+} - \frac{1}{\omega + E + \mathrm{i}0^+},$$

which has poles at $\omega = \pm E$—the excitation energy. The spectral function is $\rho(\omega) = \delta(\omega - E) - \delta(\omega + E)$: two delta-function peaks at the transition energy.

The Born rule gives the probability of finding the system in state $|1\rangle$ at time t, given initial state $|0\rangle$:

$$P_{0\to 1}(t) = |\langle 1|e^{-\mathrm{i}Ht}|0\rangle|^2 = |\langle 1|0\rangle|^2 = 0.$$

(The states $|0\rangle$ and $|1\rangle$ are eigenstates, so the transition probability is zero for a time-independent Hamiltonian.) If we add an off-diagonal perturbation $V = g(|0\rangle\langle 1| + |1\rangle\langle 0|)$, the eigenstates mix, and the transition probability oscillates: $P_{0\to 1}(t) = \frac{g^2}{E'^2}\sin^2(E't)$, where $E' = \sqrt{(E/2)^2 + g^2}$.

This simple example illustrates the complete chain: Euclidean correlator (exponential decay with rate E) → analytic continuation → retarded propagator (poles at $\pm E$) → spectral function (delta peaks) → real-time dynamics (Rabi oscillations, via the Born rule). The "gap" E that we compute throughout this book has the physical meaning of an excitation energy, and the spectral function tells us how that excitation manifests in real-time measurements.

Chapter 50

The Operator Formalism: Hamiltonians, Fock Spaces, and the HST

The path-integral formulation used throughout this book treats fields as integration variables in a functional integral. The operator formalism treats fields as operators acting on a Hilbert space. The two formulations are equivalent (connected by the transfer matrix), but each illuminates different aspects. This chapter develops the operator perspective, showing how the HST, collective modes, and quasiparticle certification appear from the Hamiltonian side.

50.1 Second Quantization and Fock Space

In the operator formalism, the fundamental field $\hat{\phi}(x)$ is an operator satisfying canonical commutation (bosonic) or anticommutation (fermionic) relations. For a bosonic field:

$$[\hat{\phi}(x), \hat{\pi}(y)] = \mathrm{i}\delta(x - y),$$

where $\hat{\pi}(x)$ is the conjugate momentum. Expanding in a complete set of single-particle states $\{u_k(x)\}$:

$$\hat{\phi}(x) = \sum_k \left(u_k(x)\hat{a}_k + u_k^*(x)\hat{a}_k^\dagger \right),$$

where $\hat{a}_k^\dagger$ and $\hat{a}_k$ are creation and annihilation operators satisfying $[\hat{a}_k, \hat{a}_{k'}^\dagger] = \delta_{kk'}$. The Fock space $\mathcal{F}$ is built by acting with creation operators on the vacuum $|0\rangle$: states $|n_{k_1}, n_{k_2}, \ldots\rangle$ with n_k quanta in mode k.

For fermions, the anticommutation relations $\{\hat{c}_k, \hat{c}_{k'}^\dagger\} = \delta_{kk'}$ enforce the Pauli exclusion principle: each mode is occupied by at most one particle.

50.2 Schwinger's Source Formalism in the Operator Language

Schwinger's approach introduces a time-dependent external source $J(x, t)$ coupled to the field operator. The Hamiltonian in the presence of the source is

$$\hat{H}[J] = \hat{H}_0 - \int \mathrm{d}^d x \, J(x, t)\hat{\phi}(x),$$

where $\hat{H}_0$ is the source-free Hamiltonian. The source J is Schwinger's probe: by varying J and measuring how the system responds, we extract all correlation functions.

The generating functional in the operator language is the vacuum persistence amplitude:

$$Z[J] = \langle 0_+|0_-\rangle_J = \langle 0|\mathcal{T}\exp\left(\mathrm{i}\int \mathrm{d}^d x\ \mathrm{d}t\, J(x,t)\hat{\phi}_I(x,t)\right)|0\rangle, \tag{50.1}$$

where $\hat{\phi}_I$ is the field in the interaction picture and $\mathcal{T}$ denotes time ordering. The connected generating functional is $W[J] = \mathrm{i}\log Z[J]$ (with a factor of i because we are in Minkowski space).

Schwinger's key insight was that the response of the system to the source encodes all physical information:

$$\frac{\delta W}{\delta J(x)} = \langle\hat{\phi}(x)\rangle_J, \qquad \frac{\delta^2 W}{\delta J(x)\delta J(y)} = \mathrm{i}\langle\mathcal{T}\hat{\phi}(x)\hat{\phi}(y)\rangle_c,$$

where $\langle\cdot\rangle_c$ denotes the connected correlator. The Legendre transform of $W[J]$ gives the effective action $\Gamma[\phi_{\mathrm{cl}}]$, whose Hessian Γ'' is the *inverse* of the connected two-point function — the precision, in the language of this book.

In the Euclidean formulation (Wick-rotating to imaginary time), this operator construction reduces to the path-integral generating functional of Chapter 1. The two are connected by the transfer matrix: the Euclidean path integral with N_τ time slices equals $\mathrm{Tr}[\hat{T}^{N_\tau}]$, where $\hat{T} = e^{-\hat{H}a}$ is the transfer matrix.

50.3 The Hamiltonian of the Joint System After HST

In the path-integral formulation, the HST introduces an auxiliary field σ and rewrites the partition function as a double integral over (ϕ,σ). What does this correspond to in the operator formalism?

Consider a fermionic system with quartic interaction $-g(\bar{\psi}\psi)^2$. After the HST, the action is

$$S[\psi,\bar{\psi},\sigma] = \int\left[\bar{\psi}(G_0^{-1}+\sigma)\psi + \frac{\sigma^2}{2g}\right].$$

The corresponding **joint Hamiltonian** in the operator formalism is

$$\hat{H}_{\mathrm{joint}} = \hat{H}_0[\hat{\psi}] + \int \mathrm{d}^d x\ \hat{\sigma}(x)\hat{\rho}(x) + \frac{1}{2g}\int \mathrm{d}^d x\ \hat{\sigma}(x)^2, \tag{50.2}$$

where $\hat{\rho}(x) = \hat{\psi}^\dagger(x)\hat{\psi}(x)$ is the density operator and $\hat{\sigma}(x)$ is now a *dynamical bosonic field* with its own canonical commutation relations $[\hat{\sigma}(x),\hat{\pi}_\sigma(y)] = \mathrm{i}\delta(x-y)$.

The **joint Fock space** is a tensor product:

$$\mathcal{F}_{\mathrm{joint}} = \mathcal{F}_{\mathrm{fermion}}\otimes\mathcal{F}_{\mathrm{boson}},$$

where $\mathcal{F}_{\mathrm{fermion}}$ is the fermionic Fock space (electron states) and $\mathcal{F}_{\mathrm{boson}}$ is the bosonic Fock space for the auxiliary field (collective-mode quanta).

However, the auxiliary field $\hat{\sigma}$ has no independent kinetic energy—its "bare" Hamiltonian is purely a mass term $\hat{\sigma}^2/(2g)$. All of its dynamics (propagation, dispersion) are *induced* by the coupling to the fermions. This is the operator-language restatement of the fact that $S_{\mathrm{eff}}[\sigma]$ has no bare gradient term—the gradient energy of σ is generated by integrating out ψ.

50.4 Blockwise Diagonalization of the Joint Hamiltonian

At the mean-field level (saddle $\bar{\sigma}$), the joint Hamiltonian (50.2) is blockwise diagonal:

$$\hat{H}_{\text{joint}} \approx \hat{H}_{\text{MF}}[\hat{\psi};\bar{\sigma}] + \hat{H}_{\text{coll}}[\hat{\eta}] + \hat{H}_{\text{coupling}}[\hat{\psi},\hat{\eta}],$$

where $\hat{\eta} = \hat{\sigma} - \bar{\sigma}$ is the fluctuation of the auxiliary field.

$\hat{H}_{\text{MF}}[\hat{\psi};\bar{\sigma}]$ is the **mean-field Hamiltonian** for the fermions in the background of $\bar{\sigma}$. Its eigenstates are the quasiparticles (Bogoliubov quasiparticles in BCS, exchange-split bands in Stoner, etc.), and its excitation energies are the eigenvalues of $M(\bar{\sigma})$.

$\hat{H}_{\text{coll}}[\hat{\eta}]$ is the **collective-mode Hamiltonian**, obtained from the Hessian K of S_{eff}. Its eigenstates are collective excitations (phonons, plasmons, magnons, Anderson–Bogoliubov modes), and its excitation energies are determined by the poles of $D = K^{-1}$.

$\hat{H}_{\text{coupling}}[\hat{\psi},\hat{\eta}]$ couples quasiparticles to collective modes. It comes from the higher-order terms ($\Gamma^{(3)}\eta^3$, etc.) in the expansion of S_{eff} and from the σ-dependence of the quasiparticle spectrum. This coupling produces:

- quasiparticle scattering off collective modes (contributing to $\operatorname{Im}\Sigma$ and the quasiparticle lifetime);
- collective-mode decay into quasiparticle pairs (Landau damping, pair-breaking continuum);
- renormalization of both quasiparticle and collective-mode energies.

The full Fock space decomposes into sectors labeled by the number of quasiparticles and collective quanta:

$$\mathcal{F}_{\text{joint}} = |0\rangle \oplus \{\text{1-qp states}\} \oplus \{\text{1-collective states}\} \oplus \{\text{2-qp states}\} \oplus \{\text{1-qp + 1-coll states}\} \oplus \cdots$$

The spectral function $\rho(\omega,k)$ of the fundamental field $\hat{\psi}$ contains contributions from each sector: a sharp quasiparticle peak from the 1-qp sector, a continuum from 2-qp and collective-decay processes, and higher-order contributions.

50.5 Quasiparticle Certification from the Operator Perspective

The question "is a quasiparticle well-defined?" has a precise answer in the operator formalism. Consider the single-particle spectral function, which is defined as

$$A(\omega,\mathbf{k}) = -\frac{1}{\pi}\operatorname{Im} G_R(\omega,\mathbf{k}),$$

where G_R is the retarded Green function. Using the Lehmann representation:

$$A(\omega,\mathbf{k}) = \sum_n \left[|\langle n|\hat{c}_{\mathbf{k}}|0\rangle|^2 \delta(\omega - E_n + E_0) + |\langle n|\hat{c}^\dagger_{\mathbf{k}}|0\rangle|^2 \delta(\omega + E_n - E_0) \right]. \tag{50.3}$$

This is a sum of delta functions at the exact excitation energies $E_n - E_0$, weighted by the overlap $|\langle n|\hat{c}_{\mathbf{k}}|0\rangle|^2$ between the one-particle-added state $\hat{c}^\dagger_{\mathbf{k}}|0\rangle$ and the exact eigenstate $|n\rangle$.

A quasiparticle is "certified" when a single eigenstate $|n_{\mathbf{k}}\rangle$ dominates the sum:

$$|\langle n_{\mathbf{k}}|\hat{c}^\dagger_{\mathbf{k}}|0\rangle|^2 = Z_{\mathbf{k}} \quad \text{(close to 1)},$$

with all other overlaps small. The quasiparticle residue $Z_{\mathbf{k}}$ measures how much of the one-particle state "survives" as a well-defined excitation. If $Z_{\mathbf{k}} \to 0$, the spectral weight is spread over many eigenstates and no quasiparticle exists.

The quasiparticle criterion from the operator side: $Z_{\mathbf{k}}$ is significantly nonzero and the corresponding eigenstate $|n_{\mathbf{k}}\rangle$ is separated from the continuum by an energy gap or by a vanishing damping rate.

For a Landau Fermi liquid at low temperature, $Z_{k_F} > 0$ (finite residue at the Fermi surface) and $\Gamma_k \propto (k - k_F)^2$ (vanishing damping rate at the Fermi surface). The quasiparticle is well-defined near k_F.

For a non-Fermi liquid or a system above its coherence temperature, $Z_{\mathbf{k}} \to 0$ or $\Gamma_{\mathbf{k}} \sim \Omega_{\mathbf{k}}$, and the spectral function is a broad, featureless continuum. No quasiparticle exists.

50.6 Hydrodynamic Modes Persist Without Quasiparticles

The most important consequence of the operator formalism for the message of this book:

> Hydrodynamic modes exist independently of quasiparticles. Even in a system where no quasiparticle is certified ($Z_{\mathbf{k}} \to 0$ for all $\mathbf{k}$), the hydrodynamic modes—sound, diffusion, thermal transport—persist as sharp, well-defined excitations at long wavelengths. This is because hydrodynamic modes are protected by conservation laws, not by the existence of long-lived single-particle excitations.

In the operator formalism, this can be seen as follows. The density-density response function $\chi_{nn}(\omega, q)$ is determined by the matrix elements $\langle n|\hat{n}_q|0\rangle$, where $\hat{n}_q = \sum_k \hat{c}^\dagger_{k+q}\hat{c}_k$ is the density operator. The conservation law $[\hat{H}, \hat{Q}] = 0$ (where $\hat{Q} = \hat{n}_{q=0}$ is the total charge) forces $\chi_{nn}(\omega, q \to 0) \to 0$ for $\omega \neq 0$. This constraint holds *regardless* of whether the intermediate states $|n\rangle$ are well-described by quasiparticles.

The quark-gluon plasma at high temperature provides a striking example: the coupling constant g_s is large enough that no quasiparticle description is valid (the spectral function of quarks and gluons is broad and featureless). Nevertheless, the plasma has well-defined sound waves, described by relativistic hydrodynamics with a specific viscosity η/s (entropy density). The sound mode is protected by conservation of energy-momentum, which holds at all coupling strengths.

Similarly, in the "strange metal" phase of high-temperature superconductors, the electronic spectral function shows no quasiparticle peak, yet the material has well-defined transport properties (resistivity $\propto T$) described by hydrodynamic or near-hydrodynamic equations.

From the path-integral perspective: the collective-field Hessian K that determines the hydrodynamic modes is a functional of the *full* microscopic propagator G (through the polarization $\Pi = \mathrm{tr}[G\Gamma G\Gamma]$), but it does not require G to have a quasiparticle form. Even if G has no pole (no quasiparticle), K can still have the hydrodynamic structure $K(\omega, q) = \chi\omega^2 - \rho_s q^2 + \cdots$ dictated by conservation laws.

50.7 Operator Derivation of the Schwinger Effective Action

The effective action $\Gamma[\phi_{\mathrm{cl}}]$ can be derived directly in the operator formalism, without reference to path integrals.

Starting from the Hamiltonian $\hat{H}$, the generating functional is the vacuum amplitude (50.1). The connected generating functional $W[J]$ satisfies

$$\frac{\delta W}{\delta J(x)} = \phi_{\text{cl}}(x) \equiv \langle \hat{\phi}(x) \rangle_J.$$

The effective action is defined as the Legendre transform $\Gamma[\phi_{\text{cl}}] = W[J_*] - J_* \cdot \phi_{\text{cl}}$, where $J_*[\phi_{\text{cl}}]$ is the source that produces the specified mean field.

The equation of motion $\frac{\delta \Gamma}{\delta \phi_{\text{cl}}} = J$ is the quantum generalization of the classical Euler–Lagrange equation. At $J = 0$, the physical vacuum satisfies $\frac{\delta \Gamma}{\delta \phi_{\text{cl}}} = 0$.

The second derivative $\Gamma''[\phi_{\text{cl}}] = G_c^{-1}$ is the inverse of the connected two-point function—the precision operator. To see this from the operator side: the connected two-point function is

$$G_c(x, y) = \langle \mathcal{T} \hat{\phi}(x) \hat{\phi}(y) \rangle - \langle \hat{\phi}(x) \rangle \langle \hat{\phi}(y) \rangle = \frac{\delta^2 W}{\delta J(x) \delta J(y)}.$$

The Legendre transform exchanges $W'' = G_c$ for $\Gamma'' = G_c^{-1}$, exactly as in finite-dimensional convex analysis. This is the covariance-precision duality that runs through the entire book, now derived from Schwinger's operator formalism.

The one-loop effective action can be obtained by expanding the Hamiltonian around the mean-field solution. Let $\hat{\phi} = \phi_{\text{cl}} + \hat{\eta}$, where $\hat{\eta}$ is the quantum fluctuation. To second order:

$$\hat{H}[\hat{\phi}] \approx H[\phi_{\text{cl}}] + \frac{1}{2} \hat{\eta} \cdot S''[\phi_{\text{cl}}] \cdot \hat{\eta} + \cdots$$

The fluctuation $\hat{\eta}$ is a collection of harmonic oscillators with frequencies determined by the eigenvalues of $S''[\phi_{\text{cl}}]$. The zero-point energy of these oscillators gives the one-loop correction:

$$\Gamma^{(\text{1-loop})}[\phi_{\text{cl}}] = S[\phi_{\text{cl}}] + \frac{1}{2} \sum_k \omega_k = S[\phi_{\text{cl}}] + \frac{1}{2} \operatorname{Tr} \sqrt{S''[\phi_{\text{cl}}]},$$

where $\omega_k = \sqrt{\lambda_k}$ and λ_k are the eigenvalues of S''. This is related to the $\operatorname{Tr} \log$ by $\frac{1}{2} \sum_k \log \omega_k = \frac{1}{4} \operatorname{Tr} \log S''$ (after regularization), recovering the path-integral result.

The operator derivation makes clear what the $\operatorname{Tr} \log$ correction physically represents: it is the sum of zero-point energies of all the fluctuation modes around the mean-field background. This is a purely quantum effect—it vanishes in the classical limit $\hbar \to 0$—and it is the same zero-point energy that produces the Casimir effect, the Lamb shift, and vacuum fluctuation phenomena.

Epilogue

The reader who has followed this book from the beginning has seen a single logical thread running through a remarkable range of physics:

A non-Gaussian probability distribution is augmented with latent variables. The microscopic variables, now conditionally Gaussian, are integrated out exactly, producing a $\mathrm{Tr} \log$ in the effective action. The saddle of this effective action determines a gap equation—a self-consistency condition involving a dressed propagator. The Hessian at the saddle gives the collective-field precision; its inversion gives the collective covariance. The microscopic contribution to the Hessian is a polarization bubble—the Gaussian covariance of the composite observable coupled to the latent field. The Neumann expansion of the inverse Hessian is the Dyson series. The quadratic approximation that produces this Hessian is the RPA.

The same saddle can gap one covariance (microscopic quasiparticles) while leaving another gapless (collective phase mode). Symmetry breaking produces Goldstone modes; conservation laws produce hydrodynamic modes. These are not exotic phenomena—they are properties of the eigenvalue spectrum of a Hessian, which is a real symmetric matrix.

The computational machinery—marginalize, differentiate, invert, take a log-determinant, perform a local Gaussian approximation—is the same machinery used throughout probability, statistics, and machine learning. The physics vocabulary—propagator, self-energy, Dyson equation, polarization, RPA, gap, Goldstone mode, Ward identity, renormalization—names specific instances of this machinery, applied to specific physical contexts.

The mathematics is cleaner than the traditional vocabulary initially suggests. Once the mathematical objects are understood, the traditional physical stories become easier to enjoy, evaluate, and use without confusing them with derivations.

Notation and Conventions

Fourier Transforms

We use the convention

$$\tilde{f}(k) = \int \mathrm{d}^d x \, e^{-\mathrm{i}k\cdot x} f(x), \qquad f(x) = \int \frac{\mathrm{d}^d k}{(2\pi)^d} \, e^{\mathrm{i}k\cdot x} \tilde{f}(k).$$

The shorthand $\int_k \equiv \int \frac{\mathrm{d}^d k}{(2\pi)^d}$.

Matsubara Frequencies

At temperature $T = 1/\beta$, Euclidean time is periodic with period β. Fields are expanded in Matsubara modes:

Bosonic: $\omega_n = 2n\pi T$, $n \in \mathbb{Z}$. Fermionic: $\omega_n = (2n+1)\pi T$, $n \in \mathbb{Z}$.

The combined momentum-frequency integral is $\int_k \equiv T \sum_{\omega_n} \int \frac{\mathrm{d}^d k}{(2\pi)^d}$.

Green Functions and Propagators

Name	Definition	Notation
Euclidean propagator	$\langle \phi(\tau, x)\phi(0, y)\rangle$	$G_E(\tau, x-y)$
Time-ordered (Feynman)	$\langle \mathcal{T}\phi(t, x)\phi(0, y)\rangle$	$G_F(t, x-y)$
Retarded	$-\mathrm{i}\theta(t)\langle[\phi(t, x), \phi(0, y)]\rangle$	$G_R(t, x-y)$
Advanced	$+\mathrm{i}\theta(-t)\langle[\phi(t, x), \phi(0, y)]\rangle$	$G_A(t, x-y)$
Spectral function	$-\frac{1}{\pi}\mathrm{Im}\, G_R(\omega, k)$	$\rho(\omega, k)$

Analytic continuation: $G_E(\mathrm{i}\omega_n, k)\big|_{\mathrm{i}\omega_n \to \omega + \mathrm{i}0^+} = G_R(\omega, k)$.

Self-Energy and Polarization Signs

Throughout this book, we define the self-energy with the convention

$$G^{-1} = G_0^{-1} + \Sigma,$$

so Σ is a **correction to the precision** (inverse covariance). Some references use $G^{-1} = G_0^{-1} - \Sigma$; their Σ differs from ours by a sign.

For the collective-field Hessian:

$$K = D_0^{-1} + \Pi,$$

where Π is the polarization (the contribution from the $\mathrm{Tr}\log$ Hessian). The collective propagator is $D = K^{-1} = (D_0^{-1} + \Pi)^{-1}$.

In some references, the dielectric function is defined as $\varepsilon = 1 - V\chi_0$ (with a minus sign and the irreducible susceptibility χ_0), while we write $\varepsilon = 1 + V\Pi$ with the corresponding sign absorbed into Π. The physics is the same; only the sign convention for the response function differs.

Coupling Constants

For ϕ^4 theory: $S \supset \frac{\lambda}{4!}\phi^4$. The HST auxiliary field couples to ϕ^2 with effective coupling $g = \lambda/2$ (for the $N = 1$ scalar) or $g = \lambda/(2N)$ (for the $O(N)$ model with $S \supset \frac{\lambda}{4N}(\phi^2)^2$).

For BCS: $S \supset -g\bar{\psi}_\uparrow\bar{\psi}_\downarrow\psi_\downarrow\psi_\uparrow$ with $g > 0$ (attractive).

Traces

Tr denotes a full trace over all indices (spacetime, momentum, spin, Nambu, color, etc.). tr denotes a trace over internal indices only (spin, Nambu, color) at a fixed spacetime or momentum point.

Consolidated Translation Dictionary

Probability, statistics, and linear algebra	Physics terminology
Random field	Euclidean field
Negative log-density	Action
Normalizing constant	Partition function
Log-normalizer	Free energy / connected generating functional
Covariance	Propagator
Precision (inverse covariance)	Inverse propagator
Reference covariance	Bare propagator
Exact covariance	Full or dressed propagator
Precision difference (exact minus reference)	Self-energy
Latent-variable augmentation	Hubbard–Stratonovich transformation
Negative log marginal density	Effective action
Mode of a marginal density	Saddle / mean-field configuration
Stationarity condition for the mode	Mean-field equation / gap equation
Mass scale induced at a saddle	Gap
Hessian of marginal negative log-density	Inverse collective propagator
Inverse Hessian	Dressed collective propagator
Covariance of a composite observable	Susceptibility / polarization
Wick-factorized composite covariance	Bubble
Neumann expansion of corrected precision inverse	Dyson resummation
Quadratic (Laplace) approx. for collective fluctuations	RPA / Gaussian fluctuation theory
Null Hessian direction from symmetry	Goldstone mode
Small precision eigenvalue	Soft mode
Slow direction from a conservation law	Hydrodynamic mode
Pole of covariance	Collective excitation
Symmetry derivative identity	Ward identity
Higher derivatives of marginal neg. log-density	Effective vertices
Non-Gaussian correction to simple bubble	Vertex correction
Marginalization over fast variables	Coarse-graining
Multiscale marginalization and rescaling	Renormalization
Fixed point of the RG flow	Scale-invariant theory
Relevant direction at a fixed point	Relevant operator
Diverging inverse Hessian eigenvalue	Diverging susceptibility / critical point
Exponential decay rate of covariance	Mass / inverse correlation length

Probability, statistics, and linear algebra	Physics terminology
Spectral peak of retarded propagator	Quasiparticle or collective mode
Imaginary part of pole	Decay rate / damping
Real-time unitary evolution	Schrödinger equation
Squared amplitude of transition	Born rule probability