IV.4 Algebraic Geometry
János Kollár

1 Introduction

Succinctly put, algebraic geometry is the study of geometry using polynomials and the investigation of polynomials using geometry.

Many of us were taught the beginnings of algebraic geometry in high school, under the name "analytic geometry." When we say that $y = mx + b$ is the equation of a line L, or that $x^2 + y^2 = r^2$ describes a circle C of radius r, we establish a basic connection between geometry and algebra.

If we want to find the points where the line L and the circle C intersect, we just substitute $mx + b$ for y in the circle equation to get $x^2 + (mx + b)^2 = r^2$ and solve the resulting quadratic equation to obtain the x coordinates of the two intersection points.

This simple example encapsulates the method of algebraic geometry: a geometric problem is translated into algebra, where it is readily solvable; conversely, we get insight into algebra problems by using geometry. It is hard to guess the solutions of systems of polynomial equations, but once a corresponding geometric picture is drawn, we start to have a qualitative understanding of them. The precise quantitative answer is then provided by algebra.

2 Polynomials and Their Geometry

Polynomials are the expressions one can put together from variables and numbers by addition and multiplication. The most familiar are one-variable polynomials such as $x^3 - x + 4$, but we can use two or three variables to get, for instance, $2x^5 - 3xy^2 + y^3$ (which has degree 5 in two variables) or $x^5 - y^7 + x^2z^8 - xyz + 1$ (which has degree 10 in three variables). In general, one can use n variables, in which case they are frequently denoted by $x_1, x_2, \dots, x_n$, and we write $f(x_1, \dots, x_n)$, $f(\boldsymbol{x})$ or simply f to denote an unspecified polynomial.

Polynomials are the only functions that computers can work with. (Although your pocket calculator is likely to have a button for logarithms, it is secretly computing a polynomial whose value at a number b agrees with $\log b$ up to many decimal places.)

We can slightly rewrite the equations we gave earlier for the line L and the circle C: as $y - mx - b = 0$ and $x^2 + y^2 - r^2 = 0$. We can then describe L and C as *zero sets*: L is the zero set of $y - mx - b$ (that is, the set of all points (x, y) such that $y - mx - b = 0$) and C is the zero set of $x^2 + y^2 - r^2$.

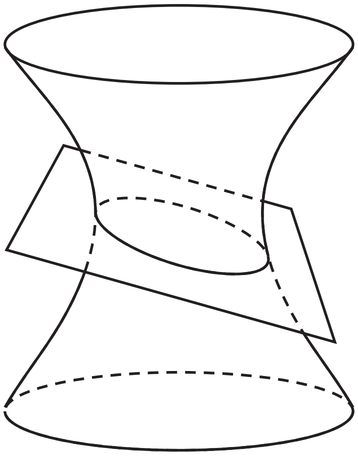

Figure 1 A hyperboloid intersecting a plane.

Similarly, the zero set of $2x^2 + 3y^2 - z^2 - 7$ in 3-space is a hyperboloid, the zero set of $z - x - y$ in 3-space is a plane, and the common zero set of these two equations in 3-space is the intersection of the hyperboloid and the plane, which is an ellipse (see figure 1).

The set of common zeros of a system of polynomial equations in any number of variables is called an *algebraic set*. These are the basic objects of algebraic geometry.

Most people feel that geometry ends in 3-space. Very few have a feeling for 4-space, also called *space-time*, and 5-space is by and large inconceivable to almost everyone. So what is the meaning of geometry in many variables?

Algebra comes to our rescue here. While I have great difficulty visualizing what a four-dimensional sphere of radius r in 5-space should be, I can easily write down its equation,

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - r^2 = 0,$$

and work with it. This equation is also something a computer can handle, which is immensely useful in applications.

I will, nonetheless, stick to two or three variables for the rest of this article. This is where all geometry starts and there are plenty of interesting questions and results.

The importance of algebraic geometry derives from the fact that significant interactions between algebra and geometry happen very frequently. Let us look at two examples, just for illustration.

3 Most Shapes Are Algebraic

Shapes that occur frequently enough to have their own name, for instance, lines, planes, circles, ellipses, hyperbolas, parabolas, hyperboloids, paraboloids, ellipsoids, are almost all algebraic. Even the more esoteric conchoid (or shell curve) of Dürer, the trident of NEWTON [VI.14], and the folium of Kepler are algebraic.

Some shapes cannot be described by polynomial equations, but they can be described by polynomial inequalities. For instance, the inequalities $0 \leqslant x \leqslant a$ and $0 \leqslant y \leqslant b$ together describe a rectangle with side lengths a, b. Shapes described by polynomial inequalities are called *semialgebraic*, and every polyhedron is semialgebraic.

Not everything is an algebraic set, though. Look, for example, at the graph of the sine function $y = \sin x$. This crosses the x-axis infinitely many times (at multiples of π). If $f(x)$ is any polynomial, then it has at most as many roots as its degree, so $y = f(x)$ will never look like $y = \sin x$.

We can, however, get very close to $\sin x$ with a polynomial if we concentrate on values of x that are not too large. For instance, the degree-7 Taylor polynomial

$$x - \tfrac{1}{6}x^3 + \tfrac{1}{120}x^5 - \tfrac{1}{5040}x^7$$

differs from $\sin x$ by an error of at most 0.1 for $-\pi < x < \pi$. This is a very special case of a basic theorem of Nash that says that every "reasonable" geometric shape is algebraic if we ignore what happens very far from the origin. So, what is reasonable? Certainly not everything. Fractals seem profoundly nonalgebraic. The nicest shapes are MANIFOLDS [I.3 §6.9], and all of these can be described by polynomials.

Nash's theorem. *Let M be any manifold in $\mathbb{R}^n$. Fix any large number R. Then there is a polynomial f whose zero set is as close to M as we want, at least inside a ball of radius R around the origin.*

4 Codes and Finite Geometries

Consider the equation $x^2 + y^2 = z^2$, which describes a double cone in 3-space (see figure 4). If we confine ourselves to natural numbers, then the solutions of $x^2 + y^2 = z^2$ are the *Pythagorean triples*, corresponding to right-angled triangles where all sides have integer lengths, of which the two best-known examples are $(3, 4, 5)$ and $(5, 12, 13)$.

Let us now look at the same equation, but declare that we care only about the *parities* of the two sides (that is, whether they are even or odd). For instance, 3^2+15^2 and 4^2 are both even, so we say that $3^2 + 15^2 \equiv 4^2 \pmod 2$. The parities of $x^2 + y^2$ and of z^2 depend only on those of x, y, and z, so we can pretend that x, y, and z are all either 0 (the even case) or 1 (the odd case). Our equation modulo 2 therefore has four solutions:

$$000,\ 011,\ 101,\ 110.$$

These look like code words in a computer message. It was quite a surprise when it was discovered that using polynomials and their solutions modulo 2 is a great—probably the best—way of constructing ERROR-CORRECTING CODES [VII.6 §§3–5].

There is something very substantial and new happening here. Let us think for a moment about what 3-space is for us. For many it is an amorphous everything, but for algebraic geometers (with DESCARTES [VI.11] as our ancestor) it is simply a collection of points described by three numbers, the x, y, and z coordinates. Let us make a jump here, and declare that "3-space modulo 2" is the collection of all "points" given by three coordinates modulo 2. Four of these are listed above, and there are four more. The beauty of algebra is that suddenly we can talk about lines, planes, spheres, cones in this "3-space having only eight points."

We do not need to stop here, and one can work modulo any integer. For example, working modulo 7, we have 0, 1, 2, 3, 4, 5, 6 as possible coordinates, and so "3-space modulo 7" has $7^3 = 343$ points.

Talking about geometry in these spaces is very intriguing, but also technically difficult. Its great reward is that one can view this process as a "discretization" of ordinary space. Working modulo n for large n (especially when n is a prime number) gets very close to the usual geometry.

This approach is especially fruitful in number-theoretic questions. It was, for instance, instrumental in Wiles's proof of Fermat's last theorem.

For more on these topics, see ARITHMETIC GEOMETRY [IV.5].

5 Snapshots of Polynomials

Consider the equation $x^2 + y^2 = R$. If $R > 0$, then the real solutions form a circle of radius $\sqrt{R}$; if $R = 0$, we get only the origin; and if $R < 0$, we get the empty set. Thus, if $R > 0$, then the geometry of the solution set determines what R is, but otherwise it does not. We can of course look at complex solutions, and the complex solutions always determine R. (For instance, the intersection points with the x-axis are $(\pm\sqrt{R}, 0)$.)

If R is a rational number, we can ask about rational solutions of $x^2 + y^2 = R$, and if R is an integer, we can also look for solutions in the "plane modulo m" for any m.

One can even look for solutions where $x = x(t)$, $y = y(t)$ are themselves polynomials in a variable t. (Most generally, we can ask for solutions where x, y are elements of any ring containing the number R.)

To my mind, the polynomial is the central object, and each time we look at solution sets we are taking a "snapshot" of the polynomial. Some snapshots are good (like the above real snapshot for $R > 0$) and some are bad (like the above real snapshot for $R < 0$).

How good can snapshots be? Can we determine a polynomial from its snapshots?

One frequently talks about "the" equation of a hyperbola, but "an" equation would be more correct. Indeed, the hyperbola $x^2 - y^2 - R = 0$ can also be given by an equation $cx^2 - cy^2 - cR = 0$, for any $c \neq 0$. We can also use the equation $(x^2 - y^2 - R)^2 = 0$, which we may well not recognize in its expanded form. Higher powers can also be used. What about the equation $f(x, y) = (x^2 - y^2 - R)(x^2 + y^2 + R^2) = 0$? If we look only at real solutions, this is still just the hyperbola since $x^2 + y^2 + R^2$ is always positive for x, y real. However, as with one-variable polynomials, one should look at all complex roots to understand everything. Then we see that $f(\sqrt{-1}R, 0) = 0$, but the complex point $(\sqrt{-1}R, 0)$ is not on the hyperbola $x^2 - y^2 - R = 0$. In general, as long as $R \neq 0$, we get that if f is a polynomial that has exactly the same complex roots as $x^2 - y^2 - R$, then $f(x, y) = c(x^2 - y^2 - R)^m$ for some m and $c \neq 0$.

Why is the $R = 0$ case different? The reason is that for $R \neq 0$ the polynomial $x^2 - y^2 - R$ is *irreducible* (that is, it cannot be written as the product of other polynomials), while $x^2 - y^2 = (x + y)(x - y)$ is reducible with *irreducible factors* $x + y$ and $x - y$. In the latter case one gets that if $g(x, y)$ is a polynomial that has exactly the same complex roots as $x^2 - y^2$, then $f = c \cdot (x + y)^m (x - y)^n$ for some m, n and $c \neq 0$.

The analogous question for systems of equations is answered by the fundamental theorem of algebraic geometry. It is sometimes called Hilbert's theorem on the zeros, but its German name is used most of the time. For simplicity, we state only the case of one equation.

Hilbert's Nullstellensatz. *Two complex polynomials f and g have the same complex solutions if and only if they have the same irreducible factors.*

We can do even better for polynomials with integer coefficients. For instance, $x^2 - y^2 - 1 = 0$ and $2(x^2 - y^2 - 1) = 0$ have the same solutions over the real or complex numbers, and the same solutions modulo p for any odd prime p, but they have different solutions modulo 2. The general result in this case is easy and simple.

Arithmetic Nullstellensatz. *Two polynomials with integer coefficients f and g have the same solutions modulo m for every m if and only if $f = \pm g$.*

6 Bézout's Theorem and Intersection Theory

If $h(x)$ is a polynomial of degree n, then it has n complex roots, at least when they are counted with multiplicity. What happens with a system $f(x, y) = g(x, y) = 0$? Geometrically we see two curves in the plane, so we expect that there will typically be finitely many intersection points.

If f, g are both linear, we have two lines in the plane. These usually intersect in a single point, but they can be parallel and they can coincide. The first case leads to the classical declaration that "parallel lines meet at infinity" and the definition of projective planes and PROJECTIVE SPACES [III.72]. (The introduction of projective spaces and the corresponding projective varieties is a key step in algebraic geometry. It is somewhat technical so we shall skip it here, but it is indispensable even at the most basic level.)

Next, consider two polynomials of degree 2, that is, two plane conics. Two smooth conics usually intersect in at most four points (just try this by drawing two ellipses). There are also some rather degenerate cases. Two conics may coincide, or, if they are both reducible, they can have a common line. In any case, we are ready to formulate a basic result, dating back to 1779.

Bézout's theorem. *Let $f_1(\boldsymbol{x}), \dots, f_n(\boldsymbol{x})$ be n polynomials in n variables, and for each i let d_i be the degree of f_i. Then either*

(i) *the equation(s) $f_1(\boldsymbol{x}) = \cdots = f_n(\boldsymbol{x}) = 0$ have at most $d_1 d_2 \cdots d_n$ solutions; or*

(ii) *the f_i vanish identically on an algebraic curve C, and so there is a continuous family of solutions.*

As an example, the second alternative happens for the system of equations $xz - y^2 = y^3 - z^2 = x^3 - z = 0$, which has (t, t^2, t^3) as a solution for any t. This case is actually quite rare. If we pick the coefficients of the polynomials f_i randomly, then the first alternative happens with probability 1.

Ideally, we would like to make the stronger claim that if the first alternative happens, then there are *exactly* $d_1d_2\cdots d_n$ solutions, but counted "with multiplicity." This actually works, and gives us our first example of an extremely useful feature of algebraic geometry. Even in very degenerate situations it is possible to define and count the multiplicities easily. This is frequently of great help since the typical (or "generic") cases are usually very hard to compute. To get around this problem, we can sometimes find a special, degenerate case where we know that the answer will be the same, but the computations are much easier.

There are two ways to think about multiplicity: one algebraic and one geometric. The algebraic definition is computationally very efficient, but somewhat technical. The geometric interpretation is easier to explain, so that is the one we shall give here, but it would be hard to compute with in practice.

If $\boldsymbol{x} = \boldsymbol{p}$ is an isolated solution of the equations $f_1(\boldsymbol{x}) = \cdots = f_n(\boldsymbol{x}) = 0$ *with multiplicity* m, then the perturbed system

$$f_1(\boldsymbol{x}) + \epsilon_1 = \cdots = f_n(\boldsymbol{x}) + \epsilon_n = 0$$

has exactly m solutions near $\boldsymbol{x} = \boldsymbol{p}$ for almost all small values of the ϵ_i.

Intersection theory is the branch of algebraic geometry that deals with generalizations of Bézout's theorem. Above, we looked at intersections of *hypersurfaces*—that is, of zero sets of single polynomials—but we may wish to look at intersections of more general algebraic sets. Also, even when the second alternative holds, we may want to count the number of isolated intersection points; this can be very tricky but also very useful.

7 Varieties, Schemes, Orbifolds, and Stacks

Consider the system $xz = yz = 0$ in 3-space. It consists of two pieces, the $z = 0$ plane and the $x = y = 0$ line. It is easy to see that neither the plane nor the line can be written as the union of algebraic sets (except by nitpickers who point out that the line is the union of the line itself and of any point on the line). In general, any algebraic set can be written in exactly one way as the union of smaller algebraic sets that in turn cannot be decomposed further. These basic building blocks are called *irreducible* algebraic sets or *algebraic varieties.*

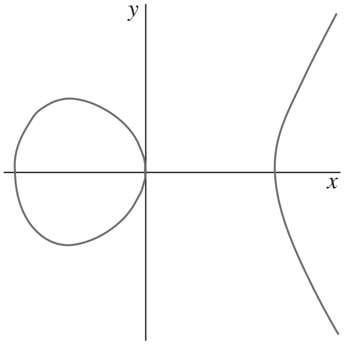

Figure 2 A smooth cubic: $y^2 = x^3 - x$.

Sometimes this is not exactly what one would naively expect. For instance, the curve in figure 2 has two connected components. The two parts are, however, not algebraic sets.

An explanation is provided by looking at the *complex* solutions of this equation. We shall see later that these form a connected set, namely a torus (with a missing point at infinity). We see two components when we look at the real solutions because we are taking a cross-section of this torus.

In general, the zero set $f = 0$ is irreducible as an algebraic set if and only if f is irreducible as a polynomial (or if it is the power of an irreducible polynomial). The implication in one direction is easy to see: if $f = gh$, then the zero set of f is the union of the zero set of g and of the zero set of h.

For many questions, keeping track only of the zero set is not enough. For instance, look at the polynomial $f = x^2(x-1)(x-2)^3$. It has degree 6 and three roots at $x = 0, 1, 2$. These roots behave differently, however, and one usually says that f has a double root at $x = 0$ and a triple root at $x = 2$. If we perturb f by adding a small number ϵ to it, then the perturbed equation $f(x) + \epsilon = 0$ has two (complex) solutions near 0, one solution near 1 and three (complex) solutions near 2. Thus, these multiplicities carry important geometric meaning about the perturbation of the equation.

Similarly, it is natural to say that while $x^2y = 0$ and $xy^3 = 0$ define the same algebraic set (consisting of the two axes), the first "assigns multiplicity 2" to the y-axis and the other "assigns multiplicity 3" to the x-axis.

More complicated things can happen for systems of equations. Consider the systems $x = y^2 = 0$ and $x^3 = y = 0$ in 3-space. Both define the z-axis and it is reasonable to say that the first does so with multiplicity 2, the second with multiplicity 3. There is, however, a further difference. In the first case the multiplicity seems to "go in the y-direction" and in the second case it seems to go in the x-direction. We can also look at other systems, like $x - cy = y^3 = 0$, if we want to see more complicated behavior.

Roughly speaking, a *scheme* is an algebraic set where we also keep track of the multiplicities and of the directions they occur in.

Consider the xy-plane and consider the map that reflects across the origin. Thus a point (x, y) is mapped to $(-x, -y)$. Let us try to glue each point (x, y) to its image $(-x, -y)$. What do we get? The right half-plane $x \geqslant 0$ is mapped to the left half-plane $x \leqslant 0$, so it is enough to work out what happens with the right half-plane. The positive y-axis is glued to the negative y-axis, and the resulting surface is a dunce cap (but less pointy).

Algebraically, it is one half of the cone $z^2 = x^2 + y^2$. This cone looks nice and smooth except at the vertex. There it is more complicated, but the above construction shows that it can be obtained from a plane by a reflection across a point. More generally, suppose we take the n-dimensional space $\mathbb{R}^n$ and finitely many symmetries of it. If we glue together points that move into each other, we again get an algebraic variety, most of whose points are smooth, but some of which are more complicated. A variety made up of pieces like these is called an *orbifold*. (When this is defined more precisely, we also keep track of which symmetries have been used.) In practice, such varieties occur frequently; that is why they deserve a separate name.

Finally, if we marry a scheme to an orbifold, the outcome is a *stack*. The study of stacks is strongly recommended to people who would have been flagellants in earlier times.

8 Curves, Surfaces, Threefolds

As with any geometric object, one of the simplest questions one can ask about a variety is: what is its dimension? As expected, a curve in the plane has dimension 1, and a surface in 3-space has dimension 2. This seems quite simple until one writes down examples like $S = (x^4 + y^4 + z^4 = 0)$, which is only the origin in $\mathbb{R}^3$. This example is, nonetheless, still two dimensional: the explanation is that we were looking at the wrong snapshot. Using complex numbers we can solve the equation as $z = \sqrt[4]{-x^4 - y^4}$, so the complex solutions of $x^4 + y^4 + z^4 = 0$ can be described by two independent variables x, y and a dependent variable z. Thus, it is quite reasonable to say that S is two dimensional.

This idea works more generally. If X is any variety in some complex space $\mathbb{C}^n$, then choose a random set of n independent directions to serve as a basis, or coordinate system, for $\mathbb{C}^n$, and hence for X. With probability 1 (i.e., except in degenerate cases) one finds that there is some d such that the first d coordinates of a point x in X can vary independently, while the rest depend on them. This number d depends on X only and is called the *dimension* (or, to be precise, the *algebraic dimension*) of X.

If X is a variety and f is a polynomial, then the intersection $X \cap (f = 0)$ has dimension one less than $\dim X$ (unless f vanishes identically on X or never takes the value zero on X).

If X is a subset of $\mathbb{R}^n$ defined by real equations, and if it is smooth (see the next section for a discussion of smoothness), then its TOPOLOGICAL DIMENSION [III.17] is the same as its algebraic dimension.

For complex varieties, the topological dimension is twice the algebraic dimension. Thus, for an algebraic geometer, $\mathbb{C}^n$ has dimension n. In particular, for us $\mathbb{C}$ is the "complex line," whereas everybody else calls this the "complex plane." Our "complex plane" is, of course, $\mathbb{C}^2$.

A variety of dimension 1 is called a *curve*. A *surface* is a variety of dimension 2, and a *threefold* is a variety of dimension 3.

The theory of algebraic curves is a very well developed and beautiful subject. We shall see later how one can start to get an overview of all algebraic curves. Surfaces have been intensively studied for the last century, and now we have reached a reasonably complete understanding of them. This is a much more complicated theory than for curves. Still very little is known for varieties of dimension 3 and up. At least conjecturally, all these dimensions behave in roughly the same way. Despite some progress, especially in dimension 3, many questions are wide open.

9 Singularities and Their Resolutions

If we look at the simplest examples of algebraic curves in figure 3, we see that most points of a curve are smooth, but that there may be a finite set of more complicated singular points. Let us compare these with the curve in figure 2.

All three curves pass through the origin, since their equation has no constant term. The equation of figure 2 has a linear term and the curve looks nice and smooth at the origin, whereas the equations of figure 3 contain

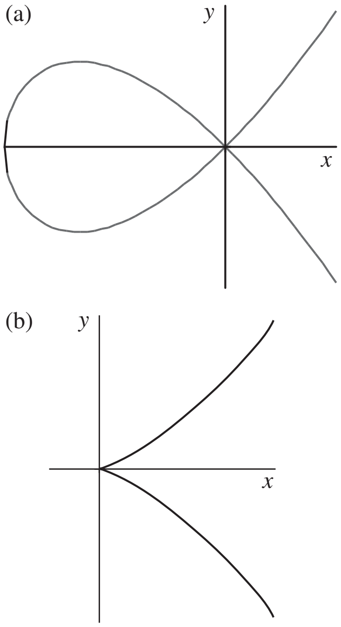

Figure 3 Singular cubics: (a) $y^2 = x^3 + x^2$ and (b) $y^2 = x^3$.

no linear term and the curves are more complicated at the origin. This is not an accident. For small values of x, the higher powers $x^2, x^3, \dots$ are much smaller than x in absolute value, so near the origin the linear terms dominate. If we have only linear terms $ax + by = 0$, we get a line through the origin, and an algebraic curve $ax + by + cx^2 + gxy + ey^2 + \cdots = 0$ is close to the line $ax + by = 0$, at least for very small values of x and y.

The study of a curve near another point with coordinates (p, q) can be reduced to the case $(p, q) = (0, 0)$ via the coordinate change $(x, y) \mapsto (x - p, y - q)$.

In general, if $f(\mathbf{0}) = 0$ and f has a (nonzero) linear term $L(f)$, the hypersurface $f = 0$ is very close to the hyperplane $L(f) = 0$. This is the so-called *implicit function theorem*. Such points are called *smooth*. Points that are not smooth are called *singular*. One can easily show that the singular points of X form an algebraic set, defined by the vanishing of all partial derivatives $\partial f / \partial x_i$. A random hypersurface will, with probability 1, be smooth, but there are many singular hypersurfaces as well.

The smooth and singular points of an arbitrary variety of dimension d can be defined analogously by comparing X with d-dimensional linear subspaces.

Singularities also occur in other geometric fields, such as topology and differential geometry, but by and large these fields shy away from their study (with the notable exception of catastrophe theory). By contrast, algebraic geometry provides very powerful tools for their investigation.

Let us start with singularities of hypersurfaces, or equivalently with *critical points* of functions. When thinking about these it is natural to work not just with polynomials but with more general power series, that is, functions $f(x_1, \dots, x_n)$ that can be written as "polynomials of infinite degree." For simplicity of notation we shall assume that $f(\mathbf{0}) = 0$. Two functions f, g are considered to be *equivalent* if there is a coordinate change $x_i \mapsto \phi_i(\boldsymbol{x})$, where each ϕ_i is given by a power series, such that $f(\phi_1(\boldsymbol{x}), \dots, \phi_n(\boldsymbol{x})) = g(\boldsymbol{x})$.

In the one-variable case, any f can be written as

$$f = x^m(a_m + a_{m+1}x + \cdots),$$

where $a_m \neq 0$. The (inverse of the) substitution

$$x \mapsto x \sqrt[m]{a_m + a_{m+1}x + \cdots}$$

then shows that f is equivalent to x^m. The functions x^m are inequivalent for different values of m, so in this particular case the lowest-degree monomial occurring in f determines f up to equivalence. (Note that even if f is a polynomial, the above change of variable involves an infinite power series: it is because we cannot invert polynomials, even locally, that it is more convenient to consider general power series.)

In general, the lowest-degree terms of a power series do not determine the singularity, but taking more terms is usually enough to do so, because of the following result.

Algebraization of analytic singularities. *Given a power series f, let $f_{\leqslant N}$ denote the polynomial obtained from f by deleting all monomials of degree greater than N. If $\mathbf{0}$ is an isolated singular point of the hypersurface $(f = 0)$, then f is equivalent to $f_{\leqslant N}$ for sufficiently large N.*

To see an example of a nonisolated singularity at $\mathbf{0}$, take

$$\begin{aligned} g(x, y, z) &= \left(y + \frac{x}{1 - x}\right)^2 - z^3 \\ &= (y + x + x^2 + x^3 + \cdots)^2 - z^3. \end{aligned}$$

It has singular points not just at $\mathbf{0}$, but everywhere along the curve $y + (x/(1 - x)) = z = 0$. On the other hand, one can easily check that all truncations $g_{\leqslant N}$ do have an isolated singular point at $\mathbf{0}$.

If we have two power series, f and g, we can view functions of the form $f + \epsilon g$ as perturbations of f. A very fruitful question of singularity theory asks:

what can we say about the perturbations of a given polynomial or power series f?

For instance, in the one-variable case, the polynomial x^m can be perturbed as $x^m + \epsilon x^r$, which is equivalent to x^r if $r < m$. Every perturbation contains x^m, so if $r > m$, then no perturbation of x^m will be equivalent to x^r (because near the origin x^m will be much larger than x^r). Hence, up to equivalence, the set of all possible perturbations of x^m is $\{x^r : r \leqslant m\}$.

On the other hand, it is not hard to see that for any given ϵ, there are only twenty-four different values of η for which the polynomials $xy(x^2-y^2)+\epsilon y^2(x^2-y^2)$ and $xy(x^2-y^2)+\eta y^2(x^2-y^2)$ are equivalent. (Indeed, both polynomials describe four lines through the origin. The first one gives the lines $y = 0$, $x = y$, $x = -y$, and $x = -\epsilon y$, and the second gives the same lines except that η replaces ϵ. The linear part of any supposed equivalence gives a linear transformation mapping the first set of four lines to the second. There are twenty-four ways to assign which line goes to which line.) Thus $xy(x^2 - y^2)$ has a continuous family of inequivalent perturbations.

Simple singularities. *Suppose that the polynomial or power series $f(x_1, \dots, x_n)$ has only finitely many inequivalent perturbations. Then f is equivalent to one of the following normal forms:*

$$\begin{array}{lll}
A_m & x_1^{m+1} + x_2^2 + \cdots + x_n^2 & (m \geqslant 1), \\
D_m & x_1^2 x_2 + x_2^{m-1} + x_3^2 + \cdots + x_n^2 & (m \geqslant 4), \\
E_6 & x_1^3 + x_2^4 + x_3^2 + \cdots + x_n^2, & \\
E_7 & x_1^3 + x_1 x_2^3 + x_3^2 + \cdots + x_n^2, & \\
E_8 & x_1^3 + x_2^5 + x_3^2 + \cdots + x_n^2. &
\end{array}$$

The names should bring to mind the CLASSIFICATION OF LIE GROUPS [III.48]. The connections are numerous but not easy to explain. When $n = 3$, these are also called *Du Val singularities* or *rational double points*.

Consider again the cone $z^2 = x^2 + y^2$. Earlier, we described a two-to-one parametrization of it. Here is another, and for many purposes better, parametrization over the real numbers.

In the (u, v, w)-space consider the smooth cylinder $u^2 + v^2 = 1$. The map $(u, v, w) \mapsto (uw, vw, w)$ maps the cylinder onto the cone (see figure 4). The map is one-to-one away from the vertex, the preimage of which is the circle $u^2 + v^2 = 1$ in the $(w = 0)$-plane.

(Sharp-eyed readers will have noticed that this map is not so nice if we use complex numbers. In general, we want parametrizations that work both for real and complex numbers, but that would be quite a bit more complicated to describe.)

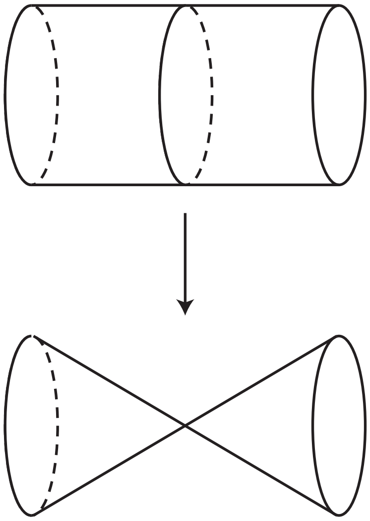

Figure 4 A resolution of the cone.

The advantage of the cylinder over the cone is that it does not have a singularity. Parametrizations of varieties in terms of smooth varieties are very useful, and there is a major result that tells us that they always exist, at least when the varieties are real or complex. (The corresponding result is still unknown for the finite geometries considered earlier.)

Resolution of singularities (Hironaka). *For any variety X there is another smooth variety Y and a polynomially defined surjective map $\pi : Y \to X$ such that π is invertible at all smooth points of X.*

(In the cone example above, one can take the whole cylinder, but the cylinder minus finitely many points in the collapsed circle would also work. In order to avoid such silly cases, we require π to be surjective in a very strong sense: if a sequence of smooth points $x_i \in X$ converges to a limit in X, then a subsequence of their preimages $\pi^{-1}(x_i)$ converges to a limit in Y.)

10 Classification of Curves

In order to get an idea of how the classification of algebraic varieties should proceed, let us look at hypersurfaces of degree d in n-space. These are given by a degree-d polynomial $f(x_1, \dots, x_n) = 0$. The set of all polynomials of degree at most d forms a vector space $V_{n,d}$. Thus hypersurfaces have two obvious discrete invariants, the dimension and the degree, and one can move between hypersurfaces of the same dimension and degree by varying the coefficients of f continuously. Moreover, the entire set $V_{n,d}$ is itself an algebraic variety. Our aim is to develop a similar understanding for all varieties, which can be done in two steps.

The first step is to define some integers, naturally attached to varieties, which stay the same if we change a variety continuously. Such integers are called *discrete invariants*. The simplest example is the dimension.

The second is to show that the set of all varieties with the same discrete invariant is parametrized by another algebraic variety, called the MODULI SPACE [IV.8]. Moreover, we would like the variety used for this parametrization to be chosen as economically as possible. We will look at this in more detail in the next section.

Let us see how it is accomplished for curves. Here there is only one more discrete invariant besides the dimension, known as the *genus* of the curve. This has many different definitions: one of the simplest is through topology. Let E be a smooth curve and let us look at its complex points. Locally, this set looks like $\mathbb{C}$, so it is a topological surface. After patching up some holes at infinity, we get a compact surface. Multiplication by $\sqrt{-1}$ gives an orientation, so basic topology tells us that we get a sphere with a certain number of handles attached (see DIFFERENTIAL TOPOLOGY [IV.7]). The genus of the curve is defined to be the number of these handles (that is, the genus of the corresponding surface). To see what this means in practice, let us look at some examples.

A line in 2-space is like the complex numbers, which can be viewed as a sphere minus a point. This sphere, $\mathbb{C}$ plus the point at infinity, is also called the *Riemann sphere*. So the genus is zero.

Next, we look at conics. Here it is better to use some projective geometry. Take any tangent of the conic and move this so that it becomes the line at infinity. Then we get a parabola, which, in suitable coordinates, is given by an equation $y = x^2$. The polynomial map $t \mapsto (t, t^2)$, with its inverse $(x, y) \mapsto x$, shows that this parabola is isomorphic to a line, so again has genus 0.

Cubics are quite a bit more complicated. A first warning is that $y = x^3$ is the wrong cubic to look at. It is smooth (and has genus 0) but it is singular at infinity. (The earlier expediency of keeping silent about projective geometry starts to bite us!) In any case, the correct thing to do is to choose the tangent line of the cubic at an inflection point and move that to infinity. After some computation we obtain a much-simplified equation $y^2 = f(x)$, where f has degree 3. What is the genus?

Consider the special case $y^2 = x(x-1)(x-2)$. We try to understand the two-to-one projection to the (complex) x-axis, but it is better to do this when the x-axis has already had the point at infinity added, so that it is the Riemann sphere. If we remove the interval $0 \leqslant x \leqslant 1$ and the half line $2 \leqslant x \leqslant +\infty$ from the Riemann sphere, then the function $y = \sqrt{x(x-1)(x-2)}$ has two branches. (This means that y takes two different values for each x, the positive and negative square roots of $x(x-1)(x-2)$, but if one moves x about, one can let y vary in a continuous way.) The sphere minus two slits is topologically like a cylinder, hence the complex cubic is glued together from two cylinders. So we get the torus and the genus is 1.

It turns out that a smooth plane curve of degree d has genus $\frac{1}{2}(d-1)(d-2)$, but I find this hard to see directly topologically.

It is a (probably hopeless) dream of algebraic geometers to give a similarly simple description of the discrete invariants for higher-dimensional varieties. Unfortunately, the topological invariants of the complex points are not good enough, and they probably mislead more than help.

As a further illustration of the approach to the classification of curves, here is a list of all curves of low genus.

Genus 0. There is only one curve of genus 0. As we saw, it can be realized as a line or as a conic in the plane.

Genus 1. Every curve of genus 1 is a plane cubic, and it can be given by an equation of the form $y^2 = f(x)$, where f has degree 3. Genus-1 curves are usually called ELLIPTIC CURVES [III.21], since they first appeared (in the guise of elliptic integrals) in connection with the arc length of ellipses. We look at these in more detail later.

Genus 2. Every curve of genus 2 can be given by an equation of the form $y^2 = f(x)$, where f has degree 5. (These curves are singular at infinity.) More generally, if f has degree $2g + 1$ or $2g + 2$, then the curve $y^2 = f(x)$ has genus g. For $g \geqslant 3$, such curves, called *hyperelliptic*, are rather special.

Genus 3. Every curve of genus 3 can be realized as a plane curve of degree 4 (or it is hyperelliptic).

Genus 4. Every curve of genus 4 can be presented as a space curve given by two equations of degrees 2 and 3 (or it is hyperelliptic).

It should be emphasized that hyperelliptic curves do not form a separate family. One can move continuously from any hyperelliptic curve to a general curve of the kind described above. This can be seen through more-complicated representations.

One can continue in this manner a bit longer, up to about genus 10, but no such explicit construction is possible when the genus is large.

11 Moduli Spaces

Let us go back to plane cubics, which we parametrized by the vector space $V_{2,3}$ of degree-3 polynomials in two variables. This is not very economical. For instance, $x^3 + 2y^3 + 1$ and $3x^3 + 6y^3 + 3$ are different polynomials, but define the same curve. Furthermore, there is not much reason to distinguish $x^3 + 2y^3 + 1$ from $2x^3 + y^3 + 1$, since they are obtained from each other by switching the two coordinate axes. More generally, as we have seen in the previous section, any cubic curve can be transformed into one given by an equation $y^2 = f(x)$, where $f = ax^3 + bx^2 + cx + d$.

This is better but not yet optimal, and there are two more steps to take. First, one can set the leading coefficient of f to be 1. Indeed, substitute $y = \sqrt{a}y_1$ and then divide the whole equation by a to get $y_1^2 = x^3 + \cdots$. Second, we can make a substitution $x = ux_1 + v$ to get another elliptic curve with equation $y^2 = f(ux_1 + v) = f_1(x_1)$, where f_1 is easy to write down explicitly. One can see that these are the only coordinate changes that we can make without messing up the form $y^2 =$ (cubic polynomial).

It is still not very clear what happens. To get a better answer, look at the three roots of f, so $f(x) = (x - r_1)(x - r_2)(x - r_3)$. (Again, complex numbers inevitably appear.) If we make the substitution $x \mapsto (r_2 - r_1)x + r_1$, we get a new polynomial $f_1(x)$, two of whose roots are 0 and 1. Thus our elliptic curve is transformed into $y^2 = x(x - 1)(x - \lambda)$. So instead of the four unknown coefficients of f, we are down to only one unknown, λ.

This form is still not completely unique. In our transformation we sent r_1, r_2 to 0, 1, but we could have used any two roots. For instance, we can substitute $x \mapsto 1-x$, sending $\lambda \mapsto 1 - \lambda$, or $x \mapsto \lambda x$, sending $\lambda \mapsto \lambda^{-1}$. All together, the six values

$$\lambda,\ \frac{1}{\lambda},\ 1-\lambda,\ \frac{1}{1-\lambda},\ \frac{-\lambda}{1-\lambda},\ \frac{1-\lambda}{-\lambda}$$

give "the same" elliptic curve. Most of the time these six values are different, but there may be coincidences. For instance, we get only three different values if $\lambda = -1$. This corresponds to the fact that the elliptic curve $y^2 = x(x - 1)(x + 1)$ has four symmetries: $(x, y) \mapsto (-x, \pm\sqrt{-1}y)$ and $(x, y) \mapsto (x, \pm y)$. (An unusual feature of elliptic curves is that they all have the second pair of symmetries. At $\lambda = 1$ we pick up 4/2 new symmetries, which corresponds to halving the number of different values above.)

The best way to think about it is to view this as an action of the symmetric group S_3 (the group of permutations of a three-element set) on the set $\mathbb{C} \setminus \{0, 1\}$.

It is not at all obvious that we have run out of tricks, but we have in fact reached the final result.

Moduli of elliptic curves. *The set of all elliptic curves is in a natural one-to-one correspondence with the points of the quotient orbifold* $(\mathbb{C} \setminus \{0, 1\})/S_3$. *The orbifold points correspond to the elliptic curves with extra automorphisms.*

This is the simplest illustration of a general phenomenon.

Moduli principle. *In most cases of interest, the set of all algebraic varieties with fixed discrete invariants is in a natural one-to-one correspondence with the points of an orbifold. The orbifold points correspond to the varieties with extra automorphisms.*

The moduli orbifold (also called the moduli space) of smooth curves of genus g is denoted by $\mathcal{M}_g$. These are among the most intensely studied orbifolds in algebraic geometry, especially since the recent discovery of their fundamental position in STRING THEORY [IV.17 §2] and MIRROR SYMMETRY [IV.16].

12 Effective Nullstellensatz

In order to show that there are still interesting elementary questions in algebraic geometry, let us try to decide when m given polynomials $f_1, \dots, f_m$ have no common complex zero. The classical answer is given by the following result, which tells us that an obviously necessary condition is in fact sufficient.

Weak Nullstellensatz. *The polynomials* $f_1, \dots, f_m$ *have no common complex zero if and only if there are polynomials* $g_1, \dots, g_m$ *such that*

$$g_1 f_1 + \cdots + g_m f_m = 1.$$

Let us now make a guess that we can find g_j with degree at most 100. We can then write

$$g_j = \sum_{i_1+\cdots+i_n \leqslant 100} a_{j,i_1,\dots,i_n} x_1^{i_1} \cdots x_n^{i_n},$$

where the $a_{j,i_1,\dots,i_n}$ are indeterminates. If we write $g_1 f_1 + \cdots + g_m f_m$ as a polynomial in the variables $x_1, \dots, x_n$, then all the coefficients must vanish, save the constant term which must equal 1. Thus we get a system of *linear* equations in the indeterminates $a_{j,i_1,\dots,i_n}$. The solvability of systems of linear equations is well-known (with good computer implementations). Thus we can decide if there is a solution with $\deg g_j \leqslant 100$. Of course it is possible that 100 was too small a guess, and we may have to repeat the process with larger and larger degree bounds. Will this ever end? The answer is given by the following result, which was proved only recently.

Effective Nullstellensatz. *Let $f_1, \dots, f_m$ be polynomials of degree less than or equal to d in n variables, where $d \geqslant 3$, $n \geqslant 2$. If they have no common zero, then $g_1 f_1 + \cdots + g_m f_m = 1$ has a solution such that $\deg g_j \leqslant d^n - d$.*

For most systems, one can find solutions such that $\deg g_j \leqslant (n-1)(d-1)$, but in general the upper bound $d^n - d$ cannot be improved.

As explained above, this provides a computational method for deciding whether or not a system of polynomial equations has a common solution. Unfortunately, this is rather useless in practice as we end up with exceedingly large linear systems. We still do not have a computationally effective and foolproof method.

13 So, What Is Algebraic Geometry?

To me algebraic geometry is a belief in the unity of geometry and algebra. The most exciting and profound developments arise from the discovery of new connections. We have seen hints of some of these; many more were left unmentioned. Born with Cartesian coordinates, algebraic geometry is now intertwined with coding theory, number theory, computer-aided geometric design, and theoretical physics. Several of these connections have emerged in the last decade, and I hope to see many more in the future.

Further Reading

Most of the algebraic geometry literature is very technical. A notable exception is *Plane Algebraic Curves* (Birkhäuser, Boston, MA, 1986), by E. Brieskorn and H. Knörrer, which starts with a long overview of algebraic curves through arts and sciences since antiquity, with many nice pictures and reproductions. *A Scrapbook of Complex Curve Theory* (American Mathematical Society, Providence, RI, 2003), by C. H. Clemens, and *Complex Algebraic Curves* (Cambridge University Press, Cambridge, 1992), by F. Kirwan, also start at an easily accessible level, but then delve more quickly into advanced subjects.

The best introduction to the techniques of algebraic geometry is *Undergraduate Algebraic Geometry* (Cambridge University Press, Cambridge, 1988), by M. Reid. For those wishing for a general overview, *An Invitation to Algebraic Geometry* (Springer, New York, 2000), by K. E. Smith, L. Kahanpää, P. Kekäläinen, and W. Traves, is a good choice, while *Algebraic Geometry* (Springer, New York, 1995), by J. Harris, and *Basic Algebraic Geometry*, volumes I and II (Springer, New York, 1994), by I. R. Shafarevich, are suitable for more systematic readings.

IV.5 Arithmetic Geometry

Jordan S. Ellenberg

1 Diophantine Problems, Alone and in Teams

Our goal is to sketch some of the essential ideas of arithmetic geometry; we begin with a problem which, on the face of it, involves no geometry and only a bit of arithmetic.

Problem. Show that the equation

$$x^2 + y^2 = 7z^2 \tag{1}$$

has no solution in nonzero rational numbers x, y, z.

(Note that it is only in the coefficient 7 that (1) differs from the Pythagorean equation $x^2 + y^2 = z^2$, which we know has *infinitely* many solutions. It is a feature of arithmetic geometry that modest changes of this kind can have drastic effects!)

Solution. Suppose x, y, z are rational numbers satisfying (1); we will derive from this a contradiction.

If n is the least common denominator of x, y, z, we can write

$$x = a/n, \quad y = b/n, \quad z = c/n$$

such that a, b, c, and n are integers. Our original equation (1) now becomes

$$\left(\frac{a}{n}\right)^2 + \left(\frac{b}{n}\right)^2 = 7\left(\frac{c}{n}\right)^2,$$

and multiplying through by n^2 one has

$$a^2 + b^2 = 7c^2. \tag{2}$$

If a, b, and c have a common factor m, then we can replace them by a/m, b/m, and c/m, and (2) still holds for these new numbers. We may therefore suppose that a, b, and c are integers with no common factor.

We now reduce the above equation modulo 7 (see MODULAR ARITHMETIC [III.58]). Denote by $\bar{a}$ and $\bar{b}$ the reductions of a and b modulo 7. The right-hand side of (2) is a multiple of 7, so it reduces to 0. We are left with

$$\bar{a}^2 + \bar{b}^2 = 0. \tag{3}$$

Now there are only seven possibilities for $\bar{a}$, and seven possibilities for $\bar{b}$. So the analysis of the solutions of (3) amounts to checking the forty-nine choices of $\bar{a}$, $\bar{b}$ and seeing which ones satisfy the equation. A few minutes of calculation are enough to convince us that (3) is satisfied only if $\bar{a} = \bar{b} = 0$.

But saying that $\bar{a} = \bar{b} = 0$ is the same as saying that a and b are both multiples of 7. This being the case, a^2 and b^2 are both multiples of 49. It follows that their sum, $7c^2$, is a multiple of 49 as well. Therefore, c^2 is a multiple of 7, and this implies that c itself is a multiple of 7. In particular, a, b, and c share a common factor of 7. We have now arrived at the desired contradiction, since we chose a, b, and c to have no common factor. Thus, the hypothesized solution leads us to a contradiction, so we are forced to conclude that there is not, in fact, any solution to (1) consisting of nonzero rational numbers.[1]

In general, the determination of rational solutions to a polynomial equation like (2) is called a *Diophantine problem*. We were able to dispose of (2) in a paragraph, but that turns out to be the exception: in general, Diophantine problems can be extraordinarily difficult. For instance, we might modify the exponents in (2) and consider the equation

$$x^5 + y^5 = 7z^5. \tag{4}$$

I do not know whether (4) has any solutions in nonzero rational numbers or not; one can be sure, though, that determining the answer would be a substantial piece of work, and it is quite possible that the most powerful techniques available to us are insufficient to answer this simple question.

More generally, one can take an arbitrary commutative RING [III.81] R, and ask whether a certain polynomial equation has solutions in R. For instance, does (2) have a solution with x, y, z in the polynomial ring $\mathbb{C}[t]$? (The answer is yes. We leave it as an exercise to find some solutions.) We call the problem of solving a polynomial equation over R a *Diophantine problem over R*. The subject of arithmetic geometry has no precise boundary, but to a first approximation one may say that it concerns the solution of Diophantine problems over subrings of NUMBER FIELDS [III.63]. (To be honest, a problem is usually called Diophantine *only* when R is a subring of a number field. However, the more general definition suits our current purposes.)

With any particular equation like (2), one can associate *infinitely many* Diophantine problems, one for each commutative ring R. A central insight—in some sense the basic insight—of modern algebraic geometry is that this whole gigantic ensemble of problems can be treated as a single entity. This widening of scope reveals structure that is invisible if we consider each problem on its own. The aggregate we make of all these Diophantine problems is called a *scheme*. We will return to schemes later, and will try, without giving precise definitions, to convey some sense of what is meant by this not very suggestive term.

A word of apology: I will give only the barest sketch of the immense progress that has taken place in arithmetic geometry in recent decades—there is simply too much to cover in an article of the present scope. I have chosen instead to discuss at some length the idea of a scheme, assuming, I hope, minimal technical knowledge on the part of the reader. In the final section, I shall discuss some outstanding problems in arithmetic geometry with the help of the ideas developed in the body of the article. It must be conceded that the theory of schemes, developed by Grothendieck and his collaborators in the 1960s, belongs to algebraic geometry as a whole, and not to arithmetic geometry alone. I think, though, that in the arithmetic setting, the use of schemes, and the concomitant extension of geometric ideas to contexts that seem "nongeometric" at first glance, is particularly central.

2 Geometry without Geometry

Before we dive into the abstract theory of schemes, let us splash around a little longer among the polynomial equations of degree 2. Though it is not obvious from our discussion so far, the solution of Diophantine problems is properly classified as part of geometry. Our goal here will be to explain why this is so.

Suppose we consider the equation

$$x^2 + y^2 = 1. \tag{5}$$

[1] Exercise: why does our argument not obtain a contradiction from the solution $x = y = z = 0$?

One can ask: which values of $x, y \in \mathbb{Q}$ satisfy (5)? This problem has a flavor very different from that of the previous section. There we looked at an equation with *no* rational solutions. We shall see in a moment that (5), by contrast, has *infinitely* many rational solutions. The solutions $x = 0$, $y = 1$ and $x = \frac{3}{5}$, $y = -\frac{4}{5}$ are representative examples. (The four solutions $(\pm 1, 0)$ and $(0, \pm 1)$ are the ones that would be said, in the usual mathematical parlance, to be "staring you in the face.")

Equation (5) is, of course, immediately recognizable as "the equation of a circle." What, precisely, do we mean by that assertion? We mean that the set of pairs of real numbers (x, y) satisfying (5) forms a circle when plotted in the Cartesian plane.

So geometry, as usually construed, makes its entrance in the figure of the circle. Now suppose that we want to find more solutions to (5). One way to proceed is as follows. Let P be the point $(1, 0)$, and let L be a line through P of slope m. Then we have the following geometric fact.

(G) The intersection of a line with a circle consists of either zero, one, or two points; the case of a single point occurs only when the line is tangent to the circle.

From (G) we conclude that, unless L is the tangent line to the circle at P, there is exactly one point other than P where the line intersects the circle. In order to find solutions (x, y) to (5), we must determine coordinates for this point. So suppose L is the line through $(1, 0)$ with slope m, which is to say it is the line L_m whose equation is $y = m(x - 1)$. Then in order to find the x-coordinates of the points of intersection between L_m and the circle, we need to solve the simultaneous equations $y = m(x - 1)$ and $x^2 + y^2 = 1$; that is, we need to solve $x^2 + m^2(x - 1)^2 = 1$ or, equivalently,

$$(1 + m^2)x^2 - 2m^2x + (m^2 - 1) = 0. \qquad (6)$$

Of course, (6) has the solution $x = 1$. How many other solutions are there? The geometric argument above leads us to believe that there is at most one solution to (6). Alternatively, we can use the following algebraic fact, which is analogous[2] to the geometric fact (G).

(A) The equation $(1 + m^2)x^2 - 2m^2x + (m^2 - 1) = 0$ has either zero, one, or two solutions in x.

Of course, the conclusion of statement (A) holds for *any* nontrivial quadratic equation in x, not just (6); it is a consequence of the factor theorem.

In this case, it is not really necessary to appeal to any theorem; one can find by direct computation that the solutions of (6) are $x = 1$ and $x = (m^2 - 1)/(m^2 + 1)$. We conclude that the intersection between the unit circle and L_m consists of $(1, 0)$ and the point P_m with coordinates

$$\left(\frac{m^2 - 1}{m^2 + 1}, \frac{-2m}{m^2 + 1}\right). \qquad (7)$$

Equation (7) establishes a correspondence $m \mapsto \mathrm{P}_m$, which associates with each slope m a solution P_m to (5). What is more, since every point on the circle, other than $(1, 0)$ itself, is joined to $(1, 0)$ by a unique line, we find that we have established a one-to-one correspondence between slopes m and solutions, other than $(1, 0)$, to equation (5).

A very nice feature of this construction is that it allows us to construct solutions to (5) not only over $\mathbb{R}$ but over smaller fields, like $\mathbb{Q}$: it is evident that, when m is rational, so are the coordinates of the solution yielded by (7). For example, taking $m = 2$ yields the solution $(\frac{3}{5}, -\frac{4}{5})$. In fact, not only does (7) show us that (5) admits infinitely many solutions over $\mathbb{Q}$, it also gives us an explicit way to *parametrize* the solutions in terms of a variable m. We leave it as an exercise to prove that the solutions of (5) over $\mathbb{Q}$, apart from $(1, 0)$, are in one-to-one correspondence with rational values of m. Alas, rare is the Diophantine problem whose solutions can be parametrized in this way! Still, polynomial equations like (5) with solutions that can be parametrized by one or more variables play a special role in arithmetic geometry; they are called *rational varieties* and constitute by any measure the best-understood class of examples in the subject.

I want to draw your attention to one essential feature of this discussion. We relied on geometric intuition (e.g., our knowledge of facts like (G)) to give us ideas about how to construct solutions to (5). On the other hand, now that we have erected an algebraic justification for our construction, we can kick away our geometric intuition as needless scaffolding. It was a geometric fact about lines and circles that *suggested* to us that (6) should have only one solution other than $x = 1$. However, once one has had that thought, one can *prove* that there is at most one such solution by means of the purely algebraic statement (A), which involves no geometry whatsoever.

[2] Note that (A), unlike (G), contains no mention of tangency; that is because the notion of tangency is more subtle in the algebraic setting, as we will see in section 4 below.

The fact that our argument can stand without any reference to geometry means that it can be applied in situations that might not, at first glance, seem geometric. For instance, suppose we wished to study solutions to (5) over the finite field $\mathbb{F}_7$. Now this solution set would not seem rightfully to be called "a circle" at all—it is just a finite set of points! Nonetheless, our geometrically inspired argument still works perfectly. The possible values of m in $\mathbb{F}_7$ are 0, 1, 2, 3, 4, 5, 6, and the corresponding solutions P_m are $(-1,0)$, $(0,-1)$, $(2,2)$, $(5,5)$, $(5,2)$, $(2,5)$, $(0,1)$. These seven points, together with $(1,0)$, form the whole solution set of (5) over $\mathbb{F}_7$.

We have now started to reap the benefits of considering a whole bundle of Diophantine problems at once; in order to find the solutions to (5) over $\mathbb{F}_7$, we used a method that was inspired by the problem of finding solutions to (5) over $\mathbb{R}$. Similarly, in general, methods suggested by geometry can help us solve Diophantine problems. And these methods, once translated into purely algebraic form, still apply in situations that do not appear to be geometric.

We must now open our minds to the possibility that the purely algebraic appearance of certain equations is deceptive. Perhaps there could be a sense of "geometry" that was general enough to include entities like the solution set of (5) over $\mathbb{F}_7$, and in which this particular example had every right to be called a "circle." And why not? It has properties a circle has: most importantly for us, it has either zero, one, or two intersection points with any line. Of course, there are features of "circleness" which this set of points lacks: infinitude, continuity, roundness, etc. But these latter qualities turn out to be inessential when we are doing arithmetic geometry. From our viewpoint the set of solutions of (5) over $\mathbb{F}_7$ has every right to be called the unit circle.

To sum up, you might think of the modern point of view as an upending of the traditional story of Cartesian space. There, we have geometric objects (curves, lines, points, surfaces) and we ask questions such as, "What is the equation of this curve?" or "What are the coordinates of that point?" The underlying object is the geometric one, and the algebra is there to tell us about its properties. For us, the situation is exactly reversed: the underlying object is the *equation*, and the various geometric properties of solution sets of the equation are merely tools that tell us about the equation's algebraic properties. For an arithmetic geometer, "the unit circle" *is* the equation $x^2 + y^2 = 1$. And the round thing on the page? That is just a *picture* of the solutions to the equation over $\mathbb{R}$. It is a distinction that makes a remarkable difference.

3 From Varieties to Rings to Schemes

In this section, we will attempt to give a clearer answer to the question, "What is a scheme?" Instead of trying to lay out a precise definition—which requires more algebraic apparatus than would fit comfortably here—we will approach the question by means of an analogy.

3.1 Adjectives and Qualities

So let us think about adjectives. Any adjective, such as "yellow" for instance, picks out a set of nouns to which the adjective applies. For each adjective A, we might call this set of nouns $\Gamma(A)$. For instance, $\Gamma(\text{"yellow"})$ is an infinite set that might look like {lemon, school bus, banana, sun, . . . }.[3] And anyone would agree that $\Gamma(A)$ is an important thing to know about A.

Now suppose that, moved by a desire for lexical parsimony, a theoretician among us suggested that adjectives could in fact be dispensed with entirely. If, instead of A, we spoke only of $\Gamma(A)$, we could get by with a grammatical theory involving only nouns.

Is this a good idea? Well, there are certainly some obvious ways that things could go wrong. For instance, what if lots of different adjectives were sent to the same set of nouns? Then our new viewpoint would be less precise than the old one. But it certainly seems that if two adjectives apply to *exactly* the same set of nouns, then it is fair to say that the adjectives are the same, or at least synonymous.

What about relationships between adjectives? For instance, we can ask of two adjectives whether one is *stronger* than another, in the way that "gigantic" is stronger than "large." Is this relationship between adjectives still visible on the level of sets of nouns? The answer is yes: it seems fair to say that A is "stronger than" B precisely when $\Gamma(A)$ is a subset of $\Gamma(B)$. In other words, what it means to say that "gigantic" is stronger than "large" is that all gigantic things are large, though some large things may not be gigantic.

So far, so good. We have paid a price in technical difficulty: it is much more cumbersome to speak of infinite sets of nouns than it was to use simple, familiar adjectives. But we have gained something, too:

[3] Of course, in real life, there are nouns whose relationship with "yellow" is not so clear-cut, but since our goal is to make this look like mathematics, let us pretend that every object in the world is either definitively yellow or definitively not yellow.

the opportunity for generalization. Our theoretician—whom we may now call a "set-theoretic grammarian"—observes that there is, perhaps, nothing special about the sets of nouns that happen to be of the form $\Gamma(A)$ for some already known adjective A. Why not take a conceptual leap and *redefine* the word "adjective" to mean "a set of nouns"? To avoid confusion with the usual meaning of "adjective," the theoretician might even use a new term, like "quality," to refer to his new objects of study.

Now we have a whole new world of qualities to play with. For example, there is a quality {"school bus", "sun"} which is stronger than "yellow," and a quality {"sun"} (not the same thing as the *noun* "sun"!) which is stronger than the qualities "yellow," "gigantic," "large," and {"school bus", "sun"}.

I may not have convinced you that, on balance, this reconception of the notion of "adjective" is a good idea. In fact, it probably is not, which is why set-theoretic grammar is not a going concern. The corresponding story in algebraic geometry, however, is quite a different matter.

3.2 Coordinate Rings

A warning: the next couple of sections will be difficult going for those not familiar with rings and ideals—such readers can either skip to section 4, or try to follow the discussion after reading RINGS, IDEALS, AND MODULES [III.81] (see also ALGEBRAIC NUMBERS [IV.1]).

Let us recall that a *complex affine variety* (from now on, just "variety") is the set of solutions over $\mathbb{C}$ to some finite set of polynomial equations. For instance, one variety V we could define is the set of points (x,y) in $\mathbb{C}^2$ satisfying our favorite equation

$$x^2 + y^2 = 1. \tag{8}$$

Then V is what we called in the previous section "the unit circle," though in fact the shape of the set of complex solutions of (8) is a sphere with two points removed. (This is not supposed to be obvious.) It is a question of general interest, given some variety X, to understand the ring of polynomial functions that take points on X to complex numbers. This ring is called the *coordinate ring* of X, and is denoted $\Gamma(X)$.

Certainly, given any polynomial in x and y, we can regard it as a function defined on our particular variety V. So is the coordinate ring of V just the polynomial ring $\mathbb{C}[x,y]$? Not quite. Consider, for instance, the function $f = 2x^2 + 2y^2 + 5$. If we evaluate this function at various points on V,

$$f(0,1) = 7,\ f(1,0) = 7,\\ f(1/\sqrt{2}, 1/\sqrt{2}) = 7,\ f(\mathrm{i}, \sqrt{2}) = 7,\ \dots,$$

we notice that f keeps taking the same value; indeed, since $x^2 + y^2 = 1$ for all $(x,y) \in V$, we see that $f = 2(x^2 + y^2) + 5$ takes the value 7 at *every* point on V. So $2x^2 + 2y^2 + 5$ and 7 are just different names for the same function on V.

So $\Gamma(V)$ is smaller than $\mathbb{C}[x,y]$; it is the ring obtained from $\mathbb{C}[x,y]$ by declaring two polynomials f and g to be the same function whenever they take the same value at every point of V. (More formally, we are defining an EQUIVALENCE RELATION [I.2 §2.3] on the set of complex polynomials in two variables.) It turns out that f and g have this property precisely when their difference is a multiple of $x^2 + y^2 - 1$. Thus, the ring of polynomial functions on V is the quotient of $\mathbb{C}[x,y]$ by the ideal generated by $x^2 + y^2 - 1$. This ring is denoted by $\mathbb{C}[x,y]/(x^2 + y^2 - 1)$.

We have shown how to attach a ring of functions to any variety. It is not hard to show that, if X and Y are two varieties, and if their coordinate rings $\Gamma(X)$ and $\Gamma(Y)$ are ISOMORPHIC [I.3 §4.1], then X and Y are in a sense the "same" variety. It is a short step from this observation to the idea of abandoning the study of varieties entirely in favor of the study of rings. Of course, we are here in the position of the set-theoretic grammarian in the parable above, with "variety" playing the part of "adjective" and "coordinate ring" the part of "set of nouns."

Happily, we can recover the geometric properties of a variety from the algebraic properties of its coordinate ring; if this were not the case, the coordinate ring would not be such a useful object! The relationship between geometry and algebra is a long story—and much of it belongs to algebraic geometry in general, not arithmetic geometry in particular—but to give the flavor, let us discuss some examples.

A straightforward geometric property of a variety is *irreducibility*. We say a variety X is *reducible* if X can be expressed as the union of two varieties X_1 and X_2, neither of which is the whole of X. For example, the variety

$$x^2 = y^2 \tag{9}$$

in $\mathbb{C}^2$ is the union of the lines $x = y$ and $x = -y$. A variety is called *irreducible* if it is not reducible. All varieties are thus built up from irreducible varieties: the relationship between irreducible varieties and general varieties is rather like the relationship between prime numbers and general positive integers.

Moving from geometry to algebra, we recall that a ring R is called an *integral domain* if, whenever f, g are nonzero elements of R, their product fg is also nonzero; the ring $\mathbb{C}[x, y]$ is a good example.

Fact. A variety X is irreducible if and only if $\Gamma(X)$ is an integral domain.

Experts will note that we are glossing over issues of "reducedness" here.

We will not prove this fact, but the following example is illustrative: consider the two functions $f = x - y$ and $g = x + y$ on the variety X defined by (9). Neither of these functions is the zero function; note, for instance, that $f(1, -1)$ is nonzero, as is $g(1, 1)$. Their product, however, is $x^2 - y^2$, which is equal to zero on X; so $\Gamma(X)$ is not an integral domain. Notice that the functions f and g that we chose are closely related to the decomposition of X as the union of two smaller varieties.

Another crucial geometric notion is that of functions from one variety to another. (It is common practice to call such functions "maps" or "morphisms"; we will use the three words interchangeably.) For instance, suppose that W is the variety in $\mathbb{C}^3$ determined by the equation $xyz = 1$. Then the map $F : \mathbb{C}^3 \to \mathbb{C}^2$ defined by

$$F(x, y, z) = \left(\frac{1}{2}(x + yz), \frac{1}{2\mathrm{i}}(x - yz)\right)$$

maps points of W to points of V.

It turns out that knowing the coordinate rings of varieties makes it very easy to see the maps between the varieties. We merely observe that if $G : V_1 \to V_2$ is a map between varieties V_1 and V_2, and if f is a polynomial function on V_2, then we have a polynomial function on V_1 that sends every point v to $f(G(v))$. This function on V_1 is denoted by $G^*(f)$. For example, if f is the function $x + y$ on V, and F is the map above, $F^*(f) = \frac{1}{2}(x + yz) + \frac{1}{2\mathrm{i}}(x - yz)$. It is easy to check that G^* is a $\mathbb{C}$-algebra homomorphism (that is, a homomorphism of rings that sends each element of $\mathbb{C}$ to itself) from $\Gamma(V_2)$ to $\Gamma(V_1)$. What is more, one has the following theorem.

Fact. For any pair of varieties V, W, the correspondence sending G to G^* is a bijection between the polynomial functions sending W to V and the $\mathbb{C}$-algebra homomorphisms from $\Gamma(V)$ to $\Gamma(W)$.

You would not be far off in thinking of the statement "there is an injective map from V to W" as analogous to "quality A is stronger than quality B."

The move to transform geometry into algebra is not something one undertakes out of sheer love of abstraction, or hatred of geometry. Instead, it is part of the universal mathematical instinct to unify seemingly disparate theories. I cannot put it any better than Dieudonné (1985) does in his *History of Algebraic Geometry*:

> ... from [the 1882 memoirs of] Kronecker and Dedekind–Weber dates the awareness of the profound analogies between algebraic geometry and the theory of algebraic numbers, which originated at the same time. Moreover, this conception of algebraic geometry is the most simple and most clear for us, trained as we are in the wielding of "abstract" algebraic notions: rings, ideals, modules, etc. But it is precisely this "abstract" character that repulsed most contemporaries, disconcerted as they were by not being able to recover the corresponding geometric notions easily. Thus the influence of the algebraic school remained very weak up until 1920. ... It certainly seems that Kronecker was the first to dream of one vast algebraico-geometric construction comprising these two theories at once; this dream has begun to be realized only recently, in our era, with the theory of schemes.

Let us therefore move on to schemes.

3.3 Schemes

We have seen that each variety X gives rise to a ring $\Gamma(X)$, and furthermore that the algebraic study of these rings can stand in for the geometric study of varieties. But just as not every set of nouns corresponds to an adjective, not every ring arises as the coordinate ring of a variety. For example, the ring $\mathbb{Z}$ of integers is not the coordinate ring of a variety, as we can see by the following argument: for every complex number a and every variety V, the constant function a is a function on V, and therefore $\mathbb{C} \subset \Gamma(V)$ for every variety V. Since $\mathbb{Z}$ does not contain $\mathbb{C}$ as a subring, it is not the coordinate ring of any variety.

Now we are ready to imitate the set-theoretic grammarian's coup de grâce. We know that some, but not all, rings arise from geometric objects (varieties); and we know that the geometry of these varieties is described by algebraic properties of these special rings. Why not, then, just consider *every* ring R to be a "geometric object" whose geometry is determined by algebraic properties of R? The grammarian needed to invent a new word, "quality," to describe his generalized adjectives; we are in the same position with our rings-that-are-not-coordinate-rings; we will call them *schemes*.

So, after all this work, the definition of scheme is rather prosaic—schemes are rings! (In fact, we are hiding some technicalities; it is correct to say that *affine schemes* are rings. Restricting our attention to affine schemes will not interfere with the phenomena that we are aiming to explain.) More interesting is to ask how we can carry out the task whose difficulty "disconcerted" the early algebraic geometers—how can we identify "geometric" features of arbitrary rings?

For instance, if R is supposed to be an arbitrary geometric object, it ought to have "points." But what are the "points" of a ring? Clearly we cannot mean by this the *elements* of the ring; for in the case $R = \Gamma(X)$, the elements of R are *functions* on X, not points on X. What we need, given a point p on X, is some entity attached to the ring R that corresponds to p.

The key observation is that we can think of p as a map from $\Gamma(X)$ to $\mathbb{C}$: given a function f from $\Gamma(X)$ we map it to the complex number $f(p)$. This map is a homomorphism, called the *evaluation homomorphism at p*. Since points on X give us homomorphisms on $\Gamma(X)$, a natural way to define the word "point" for the ring $R = \Gamma(X)$, without using geometry, is to say that a "point" is a homomorphism from R to $\mathbb{C}$. It turns out that the kernel of such a homomorphism is a maximal ideal, i.e., a proper ideal in R which is contained in no larger ideal except R itself. Moreover, every maximal ideal of R arises from a point p of X. So a very concise way to describe the points of X might be to say that they are the maximal ideals of R. A modern algebraic geometer would say that all *prime* ideals correspond to points, not only the maximal ones. The "points" corresponding to the nonmaximal ideals are not points in the usual sense of the term; for instance, the point corresponding to the zero ideal (when it is prime) is the "generic point," which is in one sense everywhere on X at once, and in another sense nowhere in particular at all. This description sounds rather woolly, but on the algebraic side the zero ideal is something quite concrete—and in fact, having a precise notion of "generic point" turns out quite often to be useful in making a certain species of vague geometric argument into a rigorous proof.

The definition we have arrived at makes sense for *all* rings R, and not just those of the form $R = \Gamma(X)$. So we might define the "points" of a ring R to be its prime ideals. The set of prime ideals of R is given the name $\operatorname{Spec} R$, and it is $\operatorname{Spec} R$ that we call the *scheme associated with R*. (More precisely, $\operatorname{Spec} R$ is defined to be a "locally ringed topological space" whose points are the prime ideals of R, but we will not need the full power of this definition for our discussion here.)

We are now in a position to elucidate our claim, made in the first section, that a scheme incorporates into one package Diophantine problems over many different rings. Suppose, for instance, that R is the ring $\mathbb{Z}[x, y]/(x^2+y^2-1)$. We are going to catalog the homomorphisms $f : R \to \mathbb{Z}$. To specify f, I merely have to tell you the values of $f(x)$ and $f(y)$ in $\mathbb{Z}$. But I cannot choose these values arbitrarily: since $x^2 + y^2 - 1 = 0$ in R, it must be the case that $f(x)^2+f(y)^2-1 = 0$ in $\mathbb{Z}$. In other words, the pair $(f(x), f(y))$ constitutes a solution over $\mathbb{Z}$ to the Diophantine equation $x^2 + y^2 = 1$. What is more, the same argument shows that, for *any* ring S, a homomorphism $f : R \to S$ yields a solution over S to $x^2 + y^2 = 1$, and vice versa. In summary,

> for each S, there is a one-to-one correspondence between the set of ring homomorphisms from R to S, and solutions over S to $x^2 + y^2 = 1$.

This behavior is what we have in mind when we say that the ring R "packages" information about Diophantine equations over different rings.

It turns out, just as one might hope, that every interesting geometric property of varieties can be computed by means of the coordinate ring, which means it can be defined not only for varieties but also for general schemes. We have already seen, for instance, that a variety X is irreducible if and only if $\Gamma(X)$ is an integral domain. Thus, we say in general that a scheme $\operatorname{Spec} R$ is irreducible if and only if R is an integral domain (or, more precisely, if the quotient of R by its nilradical is an integral domain). One can speak of the connectedness of a scheme, its dimension, whether it is smooth, and so forth. All these geometric properties turn out, like irreducibility, to have purely algebraic descriptions. In fact, to the arithmetic geometer's way of thinking, all these *are*, at bottom, algebraic properties.

3.4 Example: Spec $\mathbb{Z}$, the Number Line

The first ring we encounter in our mathematical education—and the ring that is the ultimate subject of number theory—is $\mathbb{Z}$, the ring of integers. How does it fit into our picture? The scheme $\operatorname{Spec} \mathbb{Z}$ has as its points the set of prime ideals of $\mathbb{Z}$, which come in two flavors: there are the principal ideals (p), with p a prime number; and there is the zero ideal.

We are supposed to think of $\mathbb{Z}$ as the ring of "functions" on Spec $\mathbb{Z}$. How can an integer be a function? Well, I merely need to tell you how to evaluate an integer n at a point of Spec $\mathbb{Z}$. If the point is a nonzero prime ideal (p), then the evaluation homomorphism at (p) is precisely the homomorphism whose kernel is (p); so the value of n at (p) is just the reduction of n modulo p. At the point (0), the evaluation homomorphism is the identity map $\mathbb{Z} \to \mathbb{Z}$; so the value of n at (0) is just n.

4 How Many Points Does a Circle Have?

We now return to the method of section 2, paying particular attention to the case where the equation $x^2 + y^2 = 1$ is considered over a finite field $\mathbb{F}_p$.

Let us write V for the scheme of solutions of $x^2 + y^2 = 1$. For any ring R, we will denote by $V(R)$ the set of solutions of $x^2 + y^2 = 1$.

If R is a finite field $\mathbb{F}_p$, the set $V(\mathbb{F}_p)$ is a subset of $\mathbb{F}_p^2$. In particular, it is a *finite* set. So it is natural to wonder how large this set is: in other words, how many points does a circle have?

In section 2, guided by our geometric intuition, we observed that, for every $m \in \mathbb{Q}$, the point

$$\mathrm{P}_m = \left(\frac{m^2-1}{m^2+1}, \frac{-2m}{m^2+1}\right)$$

lies on V.

The algebraic computation showing that P_m satisfies the equation $x^2 + y^2 = 1$ is no different over a finite field. So we might be inclined to think that $V(\mathbb{F}_p)$ consists of $p+1$ points: namely, the points P_m for each $m \in \mathbb{F}_p$, together with $(1,0)$.

But this is not right: for instance, when $p = 5$ it is easy to check that the four points $(0,1)$, $(0,-1)$, $(1,0)$, $(-1,0)$ make up all of $V(\mathbb{F}_5)$. Computing P_m for various m, we quickly discover the problem; when m is 2 or 3, the formula for P_m does not make sense, because the denominator m^2+1 is zero! This is a wrinkle we did not see over $\mathbb{Q}$, where m^2+1 was always positive.

What is the geometric story here? Consider the intersection of the line L_2, that is, the line $y = 2(x-1)$, with V. If (x,y) belongs to this intersection, then $x^2+(2(x-1))^2 = 1$, so $5x^2-8x+3 = 0$. Since $5 = 0$ and $8 = 3$ in $\mathbb{F}_5$, this equation can be written as $3 - 3x = 0$; in other words, $x = 1$, which in turn implies that $y = 0$. In other words, the line L_2 intersects the circle V at only one point!

We are left with two possibilities, both disturbing to our geometric intuition. We might declare that L_2 is tangent to V; but this means that V would have multiple tangents at $(1,0)$, since the vertical line $x = 1$ should surely still be considered a tangent. The alternative is to declare that L_2 is *not* tangent to V; but then we are in the equally unsavory situation of having a line which, while not tangent to the circle V, intersects it at only one point. You are now beginning to see why I did not include an algebraic definition of "tangent" in statement (A) above!

This quandary illustrates the nature of arithmetic geometry nicely. When we move into novel contexts, like geometry over $\mathbb{F}_p$, some features stay fixed (such as "a line intersects a circle in at most two points"), while others have to be discarded (such as "there exists exactly one line, which we may call the tangent line to the circle at $(1,0)$, that intersects the circle at $(1,0)$ and no other point"[4]).

Notwithstanding these subtleties, we are now ready to compute the number of points in $V(\mathbb{F}_p)$. First of all, when $p = 2$ one can check directly that $(0,1)$ and $(1,0)$ are the only two points in $V(\mathbb{F}_2)$. Having treated this case, we assume for the rest of this section that p is odd. It follows from basic number theory that the equation $m^2+1 = 0$ has a solution in $\mathbb{F}_p$ if and only if $p \equiv 1 \pmod 4$, in which case there are exactly two such m. So, if $p \equiv 3 \pmod 4$, then every line L_m intersects the circle at a point other than $(1,0)$, and we have $p+1$ points in all. If $p \equiv 1 \pmod 4$, there are two choices of m for which L_m intersects V only at $(1,0)$; eliminating these two choices of m yields a total of $p-1$ points in $V(\mathbb{F}_p)$.

We conclude that $|V(\mathbb{F}_p)|$ is equal to 2 when $p = 2$, to $p-1$ when $p \equiv 1 \pmod 4$, and to $p+1$ when $p \equiv 3 \pmod 4$. The interested reader will find the following exercises useful: how many solutions are there to $x^2 + 3y^2 = 1$ over $\mathbb{F}_p$? What about $x^2 + y^2 = 0$?

More generally, let X be the scheme of solutions of *any* system of equations

$$F_1(x_1,\dots,x_n) = 0,\ F_2(x_1,\dots,x_n) = 0,\ \dots, \qquad (10)$$

where the F_i are polynomials with integral coefficients. Then one can associate with F a list of integers $N_2(X), N_3(X), N_5(X), \dots$, where $N_p(X)$ is the number of solutions to (10) with $x_1,\dots,x_n \in \mathbb{F}_p$. This list of integers turns out to contain a surprising amount of geometric information about the scheme X; even for the simplest schemes, the analysis of these lists is a deep problem of intense current interest, as we will see in the next section.

[4] In this case, the right attitude to adopt is that L_2 is not tangent to V, but that there are certain nontangent lines that intersect the circle at a single point.

5 Some Problems in Classical and Contemporary Arithmetic Geometry

In this section I will try to give an impression of a few of arithmetic geometry's great successes, and to gesture at some problems of current interest for researchers in the area.

A word of warning is in order. In what follows, I will be trying to give brief and nontechnical descriptions of some mathematics of extreme depth and complexity. Consequently, I will feel very free to oversimplify. I will try to avoid making assertions that are actually false, but I will often use definitions (like that of the L-function attached to an elliptic curve) that do not exactly agree with those in the literature.

5.1 From Fermat to Birch–Swinnerton-Dyer

The world is not lacking in expositions of the proof of FERMAT'S LAST THEOREM [V.10] and I will not attempt to give another one here, although it is without question the most notable contemporary achievement in arithmetic geometry. (Here I am using the mathematician's sense of "contemporary," which, as the old joke goes, means "theorems proved since I entered graduate school." The shorthand for "theorems proved before I entered graduate school" is "classical.") I will content myself with making some comments about the structure of the proof, emphasizing connections with the parts of arithmetic geometry we have discussed above.

Fermat's last theorem (rightly called "Fermat's conjecture," since it is almost impossible to imagine that FERMAT [VI.12] proved it) asserts that the equation

$$A^\ell + B^\ell = C^\ell, \qquad (11)$$

where ℓ is an odd prime, has no solutions in positive integers A, B, C.

The proof uses the crucial idea, introduced independently by Frey and Hellegouarch, of associating with any solution (A, B, C) of (11) a certain variety $X_{A,B}$, namely the curve described by the equation

$$y^2 = x(x - A^\ell)(x + B^\ell).$$

What can we say about $N_p(X_{A,B})$? We begin with a simple heuristic. There are p choices for x in $\mathbb{F}_p$. For each choice of x, there are either zero, one, or two choices for y, depending on whether $x(x - A^\ell)(x + B^\ell)$ is a quadratic nonresidue, zero, or a quadratic residue in $\mathbb{F}_p$. Since there are equally many quadratic residues and nonresidues in $\mathbb{F}_p$, we might guess that those two cases arise equally often. If so, there would on average be one choice of y for each of the p choices of x, which inclines us to make the estimate $N_p(X_{A,B}) \sim p$. Define a_p to be the error in this estimate: $a_p = p - N_p(X_{A,B})$. It is worth remembering that when X was the scheme attached to $x^2 + y^2 = 1$, the behavior of $p - N_p(X)$ was very regular; in particular, this quantity took the value 1 at primes congruent to 1 mod 4 and -1 at primes congruent to 3 mod 4. (We note, in particular, that the heuristic estimate $N_p(X) \sim p$ is quite good in this case.) Might one hope that a_p displays the same kind of regularity?

In fact, the behavior of the a_p is very *irregular*, as a famous theorem of Mazur shows; not only do the a_p fail to vary periodically, even their reductions modulo various primes are irregular!

Fact (Mazur). Suppose that ℓ is a prime greater than 3, and let b be a positive integer. It is not the case that a_p takes the same value $(\bmod\, \ell)$ for all primes p congruent to 1 $(\bmod\, b)$.[5]

On the other hand—if I may compress a 200-page paper into a slogan—Wiles proved that, when A, B, C is a solution to (11), the reductions mod ℓ of the a_p *necessarily* behaved periodically, contradicting Mazur's theorem when $\ell > 3$. The case $\ell = 3$ is an old theorem of EULER [VI.19]. This completes the proof of Fermat's conjecture, and, I hope, bolsters our assertion that the careful study of the values $N_p(X)$ is an interesting way to study a variety X!

But the story does not end with Fermat. In general, if $f(x)$ is a cubic polynomial with coefficients in $\mathbb{Z}$ and no repeated roots, the curve E defined by the equation

$$y^2 = f(x) \qquad (12)$$

is called an ELLIPTIC CURVE [III.21] (note well that an elliptic curve is not an ellipse). The study of rational points on elliptic curves (that is, pairs of rational numbers satisfying (12)) has been occupying arithmetic geometers since before our subject existed as such; a decent treatment of the story would fill a book, as indeed it does fill the book of Silverman and Tate (1992). We can define $a_p(E)$ to be $p - N_p(E)$ as above. First of all, if our heuristic $N_p(E) \sim p$ is a good estimate, we might expect that $a_p(E)$ is small compared with p; and, in fact, a theorem of Hasse from the 1930s shows that $a_p(E) \leqslant 2\sqrt{p}$ for all but finitely many p.

[5] The theorem proved by Mazur is stated by him in a very different and much more general way: he proves that certain *modular curves* do not possess any rational points. This implies that a version of the fact above is true, not only for $X_{A,B}$, but for *any* equation of the form $y^2 = f(x)$, where f is a cubic polynomial without repeated roots. We will leave it to the other able treatments of Fermat to develop that point of view.

It turns out that some elliptic curves have infinitely many rational points, and some only finitely many. One might expect that an elliptic curve with many points over $\mathbb{Q}$ would tend to have more points over finite fields as well, since the coordinates of a rational point can be reduced mod p to yield a point over the finite field $\mathbb{F}_p$. Conversely, one might imagine that, by knowing the list of numbers a_p, one could draw conclusions about the points of E over $\mathbb{Q}$.

In order to draw such conclusions, one needs a nice way to package the information of the infinite list of integers a_p. Such a package is given by the L-FUNCTION [III.47] of the elliptic curve, defined to be the following function of a variable s:

$$L(E,s) = \prod\nolimits_p' (1 - a_p p^{-s} + p^{1-2s})^{-1}. \qquad (13)$$

The notation $\prod'$ means that this product is evaluated over all primes apart from a finite set, which is easy to determine from the polynomial f. (As is often the case, we are oversimplifying; what I have written here differs in some irrelevant-to-us respects from what is usually called $L(E,s)$ in the literature.) It is not hard to check that (13) is a convergent product when s is a real number greater than $\frac{3}{2}$. Not much deeper is the fact that the right-hand side of (13) is well-defined when s is a complex number whose real part exceeds $\frac{3}{2}$. What *is* much deeper—following from the theorem of Wiles, together with later theorems of Breuil, Conrad, Diamond, and Taylor—is that we can extend $L(E,s)$ to a HOLOMORPHIC FUNCTION [I.3 §5.6] defined for *every* complex number s.

A heuristic argument might suggest the following relationship between the values of $N_p(E)$ and the value of $L(E,1)$. If the a_p are typically negative (corresponding to the $N_p(E)$ typically being greater than p) the terms in the infinite product tend to be smaller than 1; when the a_p are positive, the terms in the product tend to be larger than 1. In particular, one might expect the value of $L(E,1)$ to be closer to 0 when E has many rational points. Of course, this heuristic should be taken with a healthy pinch of salt, given that $L(E,1)$ is not in fact defined by the infinite product on the right-hand side of (13)! Nonetheless, THE BIRCH–SWINNERTON-DYER CONJECTURE [V.4], which makes precise the heuristic prediction above, is widely believed, and supported by many partial results and numerical experiments. We do not have the space here to state the conjecture in full generality. However, the following conjecture would follow from Birch–Swinnerton-Dyer.

Conjecture. The elliptic curve E has infinitely many points over $\mathbb{Q}$ if and only if $L(E,1) = 0$.

Kolyvagin proved one direction of this conjecture in 1988: that E has finitely many rational points if $L(E,1) \neq 0$. (To be precise, he proved a theorem that yields the assertion here once combined with the later theorems of Wiles and others.) It follows from a theorem of Gross and Zagier that E has infinitely many rational points if $L(E,s)$ has a *simple* zero at $s = 1$. That more or less sums up our present knowledge about the relationship between L-functions and rational points on elliptic curves. This lack of knowledge has not, however, prevented us from constructing a complex of ever more rarefied conjectures in the same vein, of which the Birch–Swinnerton-Dyer conjecture is only a tiny and relatively down-to-earth sliver.

Before we leave the subject of counting points behind, we will pause and point out one more beautiful result: the theorem of ANDRÉ WEIL [VI.93] bounding the number of points on a curve over a finite field. (Because we have not introduced projective geometry, we will satisfy ourselves with a somewhat less beautiful formulation than the usual one.) Let $F(x,y)$ be an irreducible polynomial in two variables, and let X be the scheme of solutions of $F(x,y) = 0$. Then the complex points of X define a certain subset of $\mathbb{C}^2$, which we call an *algebraic curve.* Since X is obtained by imposing one polynomial condition on the points of $\mathbb{C}^2$, we expect that X has complex dimension 1, which is to say it has real dimension 2. Topologically speaking, $X(\mathbb{C})$ is, therefore, a surface. It turns out that, for almost all choices of F, the surface $X(\mathbb{C})$ will have the topology of a "g-holed doughnut" with d points removed, for some nonnegative integers g and d. In this case we say that X is a *curve of genus g*.

In section 2 we saw that the behavior of schemes over finite fields seemed to "remember" facts arising from our geometric intuition over $\mathbb{R}$ and $\mathbb{C}$: our example there was the fact that circles and lines intersect in at most two points.

The theorem of Weil reveals a similar, though much deeper, phenomenon.

Fact. Suppose the scheme X of solutions of $F(x,y)$ is a curve of genus g. Then, for all but finitely many primes p, the number of points of X over $\mathbb{F}_p$ is at most $p + 1 + 2g\sqrt{p}$ and at least $p + 1 - 2g\sqrt{p} - d$.

Weil's theorem illustrates the startlingly close bonds between geometry and arithmetic. The more complicated the topology of $X(\mathbb{C})$, the further the number of

$\mathbb{F}_p$-points can vary from the "expected" answer of p. What is more, it turns out that knowing the size of the set $X(\mathbb{F}_q)$ for every finite field $\mathbb{F}_q$ allows us to determine the genus of X. In other words, the *finite sets of points* $X(\mathbb{F}_q)$ somehow "remember" the topology of the space of complex points $X(\mathbb{C})$! In modern language, we say that there is a theory applying to general schemes, called *étale cohomology*, which mimics the theory of cohomology applying to the topology of varieties over $\mathbb{C}$.

Let us return for a moment to our favorite curve, by taking the polynomial $F(x,y) = x^2 + y^2 - 1$. In this case, it turns out that $X(\mathbb{C})$ has $g = 0$ and $d = 2$: our previous result that $X(\mathbb{F}_p)$ contains either $p+1$ or $p-1$ points therefore conforms exactly with the Weil bounds. We also remark that elliptic curves always have genus 1; so the theorem of Hasse alluded to above is a special case of Weil's theorem as well.

Recall from section 2 that the solutions to $x^2 + y^2 = 1$, over $\mathbb{R}$, over $\mathbb{Q}$, or over various finite fields, could be parametrized by the variable m. It was this parametrization that enabled us to determine a simple formula for the size of $X(\mathbb{F}_p)$ in this case. We remarked earlier that most schemes could not be so parametrized; now we can make that statement a bit more precise, at least for algebraic curves.

Fact. If X is a genus-0 curve, then the points of X can be parametrized by a single variable.

The converse of this fact is more or less true as well (though stating it properly requires us to say more than we can here about "singular curves"). In other words, a thoroughly algebraic question—whether the solutions of a Diophantine equation can be parametrized—is hereby given a geometric answer.

5.2 Rational Points on Curves

As we said above, some elliptic curves (which are curves of genus 1) have finitely many rational points, and others have infinitely many. What is the situation for algebraic curves of other flavors?

We have already encountered a curve of genus 0 with infinitely many points: namely, the curve $x^2 + y^2 = 1$. On the other hand, the curve $x^2 + y^2 = 7$ also has genus 0, and a simple modification of the argument of the first section shows that this curve has *no* rational points. It turns out these are the only two possibilities.

Fact. If X is a curve of genus 0, then $X(\mathbb{Q})$ is either empty or infinite.

Genus-1 curves are known to fall into a similar dichotomy, thanks to the theorem of Mazur we alluded to earlier.

Fact. If X is a genus-1 curve, then either X has at most sixteen rational points or it has infinitely many rational points.

What about curves of higher genus? In the early 1920s, Mordell made the following conjecture.

Conjecture. If X is a curve of genus greater than 2, then X has finitely many rational points.

This conjecture was proved by Faltings in 1983; in fact, he proved a more general theorem of which this conjecture is a special case. It is worth remarking that the work of Faltings involves a great deal of importation of geometric intuition to the study of the scheme Spec $\mathbb{Z}$.

When you prove that a set is finite, it is natural to wonder whether you can bound its size. For example, if $f(x)$ is a degree 6 polynomial with no repeated roots, the curve $y^2 = f(x)$ turns out to have genus 2; so by Faltings's theorem there are only finitely many pairs of rational numbers (x,y) satisfying $y^2 = f(x)$.

Question. Is there a constant B such that, for all degree 6 polynomials with coefficients in $\mathbb{Q}$ and no repeated roots, the equation $y^2 = f(x)$ has at most B solutions?

This question remains open, and I do not think there is a strong consensus about whether the answer will be yes or no. The current world record is held by the curve

$$y^2 = 378\,371\,081x^2(x^2-9)^2 - 229\,833\,600(x^2-1)^2,$$

which was constructed by Keller and Kulesz and has 588 rational points.

Interest in the above question comes from its relation to a conjecture of Lang, which involves points on higher-dimensional varieties. Caporaso, Harris, and Mazur showed that Lang's conjecture implies a positive answer to the question above. This suggests a natural attack on the conjecture: if one can find a way to construct an infinite sequence of degree 6 polynomials $f(x)$ so that the equations $y = f(x)$ have ever more numerous rational solutions, then one has a disproof of Lang's conjecture! No one has yet been successful at this task. If one could *prove* that the answer to the question above was affirmative, it would probably bolster our faith in the correctness of Lang's conjecture, though of course it would bring us no nearer to turning the conjecture into a theorem.

In this article we have seen only a glimpse of the modern theory of arithmetic geometry, and perhaps I have overemphasized mathematicians' successes at the expense of the much larger territory of questions, like Lang's conjecture above, about which we remain wholly ignorant. At this stage in the history of mathematics, we can confidently say that the schemes attached to Diophantine problems *have geometry*. What remains is to say as much as we can about *what this geometry is like*, and in this respect, despite the progress described here, our understanding is still quite unsatisfactory when compared with our knowledge of more classical geometric situations.

Further Reading

Dieudonné, J. 1985. *History of Algebraic Geometry*. Monterey, CA: Wadsworth.

Silverman, J., and J. Tate. 1992. *Rational Points on Elliptic Curves*. New York: Springer.

IV.6 Algebraic Topology

Burt Totaro

Introduction

Topology is concerned with the properties of a geometric shape that are unchanged when we continuously deform it. In more technical terms, topology tries to classify TOPOLOGICAL SPACES [III.90], where two spaces are considered the same if they are homeomorphic. Algebraic topology assigns numbers to a topological space, which can be thought of as the "number of holes" in that space. These holes can be used to show that two spaces are not homeomorphic: if they have different numbers of holes of some kind, then one cannot be a continuous deformation of the other. In the happiest cases, we can hope to show the converse statement: that two spaces with the same number of holes (in some precise sense) *are* homeomorphic.

Topology is a relatively new branch of mathematics, with its origins in the nineteenth century. Before that, mathematics usually sought to solve problems exactly: to solve an equation, to find the path of a falling body, to compute the probability that a game of dice will lead to bankruptcy. As the complexity of mathematical problems grew, it became clear that most problems would never be solved by an exact formula: a classic example is the problem, known as THE THREE-BODY PROBLEM [V.33], of computing the future movements of Earth, the Sun, and the Moon under the influence of gravity. Topology allows the possibility of making qualitative predictions when quantitative ones are impossible. For example, a simple topological fact is that a trip from New York to Montevideo must cross the equator at some point, although we cannot say exactly where.

1 Connectedness and Intersection Numbers

Perhaps the simplest topological property is one called *connectedness*. This can be defined in various ways, as we shall see in a moment, but once we have a notion of what it means for a space to be connected we can then divide a topological space up into connected pieces, called *components*. The number of these pieces is a simple but useful INVARIANT [I.4 §2.2]: if two spaces have different numbers of connected components, then they are not homeomorphic.

For nice topological spaces, the different definitions of connectedness are equivalent. However, they can be generalized to give ways of measuring the number of holes in a space; these generalizations are interestingly different and all of them are important.

The first interpretation of connectedness uses the notion of a *path*, which is defined to be a continuous mapping f from the unit interval $[0,1]$ to a given space X. (We think of f as a path from $f(0)$ to $f(1)$.) Let us declare two points of X to be equivalent if there is a path from one to the other. The set of EQUIVALENCE CLASSES [I.2 §2.3] is called the set of *path components* of X and is written $\pi_0(X)$. This is a very natural way of defining the "number of connected pieces" into which X breaks up. One can generalize this notion by considering mappings into X from other standard spaces such as spheres: this leads to the notion of homotopy groups, which will be the topic of section 2.

A different way of thinking about connectedness is based on functions from X to the real line rather than functions from a line segment into X. Let us assume that we are in a situation where it makes sense to differentiate functions on X. For example, X could be an open subset of some Euclidean space, or more generally a SMOOTH MANIFOLD [I.3 §6.9]. Consider all the real-valued functions on X whose derivative is everywhere equal to zero: these functions form a real VECTOR SPACE [I.3 §2.3], which we call $H^0(X,\mathbb{R})$ (the "zeroth cohomology group of X with real coefficients"). Calculus tells us that if a function defined on an interval has derivative zero, then it must be constant, but that is not true when the domain has several connected pieces: all we can say then is that the function is constant on each connected piece of X. The number of degrees of freedom of such a function is therefore equal to the number of connected pieces, so the dimension of the vector space $H^0(X, \mathbb{R})$ is another way to describe the number of connected components of X. This is the simplest example of a cohomology group. Cohomology will be discussed in section 4.

We can use the idea of connectedness to prove a serious theorem of algebra: every real polynomial of odd degree has a real root. For example, there must be some real number x such that $x^3 + 3x - 4 = 0$. The basic observation is that when x is a large positive number or a highly negative number, the term x^3 is much bigger (in absolute value) than the other terms of the polynomial. Since this top term is an odd power of x, we have $f(x) > 0$ for some positive number x and $f(x) < 0$ for some negative number x. If f were never equal to zero, then it would be a continuous mapping from the real line into the real line minus the origin. But the real line is connected, while the real line minus the origin has two connected components, the positive and negative numbers. It is easy to show that a continuous map from a connected space X to another space Y must map X into just one connected component of Y: in our case, this contradicts the fact that f takes both positive and negative values. Therefore f must be equal to zero at some point, and the proof is complete.

This argument can be phrased in terms of the "intermediate value theorem" of calculus, which is indeed one of the most basic topological theorems. An equivalent reformulation of this theorem states that a continuous curve that goes from the lower half-plane to the upper half-plane must cross the horizontal axis at some point. This idea leads to *intersection numbers*, one of the most useful concepts in topology. Let M be a smooth oriented manifold. (Roughly speaking, a manifold is oriented if you cannot continuously slide a shape about inside it and end up with a reflection of that shape. The simplest nonoriented manifold is a Möbius strip: to reflect a shape, slide it around the strip an odd number of times.) Let A and B be two closed oriented submanifolds of M with dimensions adding up to the dimension of M. Finally, suppose that A and B intersect transversely, so that their intersection has the "correct" dimension, namely 0, and is therefore a collection of separated points.

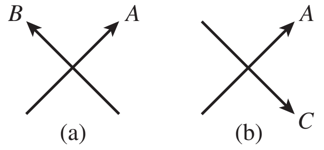

Figure 1 Intersection numbers: (a) $A \cdot B = 1$; (b) $A \cdot C = -1$.

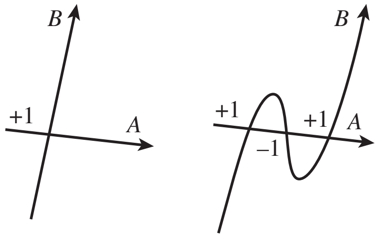

Figure 2 Moving a submanifold.

Now let p be one of these points. There is a way of assigning a weight of $+1$ or -1 to p, which depends in a natural way on the relationship between the orientations of A, B, and M (see figure 1). For example, if M is a sphere, A is the equator of M, B is a closed curve, and appropriate directions are given to A and B, then the weight of p will tell you whether B crosses A upwards or downwards at p. If A and B intersect in only finitely many points, then we can define the intersection number of A and B, written $A \cdot B$, to be the sum of the weights ($+1$ or -1) at all the intersection points. In particular, this will happen if M is COMPACT [III.9] (that is, we can think of it as a closed bounded subset of $\mathbb{R}^N$ for some N).

The important point about the intersection number is that it is an *invariant*, in the following sense: if you move A and B about in a continuous way, ending up with another pair of transverse submanifolds A' and B', then the intersection number $A' \cdot B'$ is the same as $A \cdot B$, even though the number of intersection points can change. To see why this might be true, consider again the case where A and B are curves and M is two dimensional: if A and B meet at a point with weight 1, we can wiggle one of them to turn that point into three points with weights 1, -1, and 1, but the total contribution to the intersection number is unchanged. This is illustrated in figure 2. As a result, the intersection number $A \cdot B$ is defined for *any* two submanifolds of complementary dimension: if they do not intersect transversely, one can move them until they do and use the definition we have just given.

In particular, if two submanifolds have nonzero intersection number, then they can never be moved to

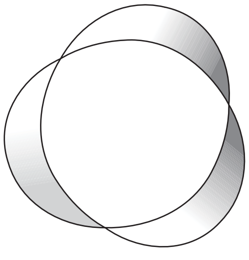

Figure 3 A surface bounded by a knot.

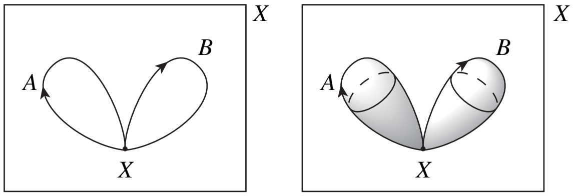

Figure 4 Multiplication in the fundamental group and in higher homotopy groups.

be disjoint from each other. This is another way to describe the earlier arguments about connectedness. It is easy to write down one curve from New York to Montevideo whose intersection number with the equator is equal to 1. Therefore, no matter how we move that curve (provided that we keep the endpoints fixed: more generally, if either A or B has a boundary, then that boundary should be kept fixed), its intersection number with the equator will always be 1, and in particular it must meet the equator in at least one point.

One of many applications of intersection numbers in topology is the idea of *linking numbers*, which comes from KNOT THEORY [III.44]. A *knot* is a path in space that begins and ends at the same point, or, more formally, a closed connected one-dimensional submanifold of $\mathbb{R}^3$. Given any knot K, it is always possible to find a surface S in $\mathbb{R}^3$ with K as its boundary (see figure 3). Now let L be a knot that is disjoint from K. The linking number of K with L is defined to be the intersection number of L with the surface S. The properties of intersection numbers imply that if the linking number of K with L is nonzero, then the knots K and L are "linked," in the sense that it is impossible to pull them apart.

2 Homotopy Groups

If we remove the origin from the plane $\mathbb{R}^2$, then we obtain a new space that is different from the plane in a fundamental way: it has a hole in it. However, we cannot detect this difference by counting components, since both the plane and the plane without the origin are connected. We begin this section by defining an invariant called the *fundamental group*, which does detect this kind of hole.

As a first approximation, one could say that the elements of the fundamental group of a space X are *loops*, which can be formally defined as continuous functions f from $[0,1]$ to X such that $f(0) = f(1)$. However, this is not quite accurate, for two reasons. The first reason, which is extremely important, is that two loops are regarded as equivalent if one can be continuously deformed to the other while all the time staying inside X. If this is the case, we say that they are *homotopic*. To be more formal about this, let us suppose that f_0 and f_1 are two loops. Then a *homotopy* between f_0 and f_1 is a collection of loops f_s in X, one for each s between 0 and 1, such that the function $F(s,t) = f_s(t)$ is a continuous function from $[0,1]^2$ to X. Thus, as s increases from 0 to 1, the loop f_s moves continuously from f_0 to f_1. If two loops are homotopic, then we count them as the same. So the elements of the homotopy group are not actually loops but equivalence classes, or *homotopy classes*, of loops.

Even this is not quite correct, because for technical reasons we need to impose an extra condition on our loops: that they all start from (and therefore end at) a given point, called the *base point*. If X is connected, it turns out not to matter what this base point is, but we need it to be the same for all loops. The reason for this is that it gives us a way to multiply two loops: if x is the base point and A and B are two loops that start and end at x, then we can define a new loop by going around A and then going around B. This is illustrated in figure 4. We regard this new loop as the product of the loops A and B. It is not hard to check that the homotopy class of this product depends only on the homotopy classes of A and B, and that the resulting binary operation turns the set of homotopy classes of loops into a GROUP [I.3 §2.1]. It is this group that we call the fundamental group of X. It is denoted $\pi_1(X)$.

The fundamental group can be computed for most of the spaces we are likely to encounter. This makes it an important way to distinguish one space from another. First of all, for any n the fundamental group of $\mathbb{R}^n$ is the trivial group with just one element, because any loop in $\mathbb{R}^n$ can be continuously shrunk to its base point. On the other hand, the fundamental group of $\mathbb{R}^2 \setminus \{0\}$, the plane with the origin removed, is isomorphic to the group $\mathbb{Z}$ of the integers. This tells us that we can associate with any loop in $\mathbb{R}^2 \setminus \{0\}$ an integer that does not change if we modify the loop in a continuous way. This integer is known as the *winding number*. Intuitively, the winding number measures the total number of times that the mapping goes around the origin, with counterclockwise circuits counting positively and clockwise ones negatively. Since the fundamental group of $\mathbb{R}^2 \setminus \{0\}$ is not the trivial group, $\mathbb{R}^2 \setminus \{0\}$ cannot be homeomorphic to the plane. (It is an interesting exercise to try to find an elementary proof of this result—that is, a proof that does not use, or implicitly reconstruct, any of the machinery of algebraic topology. Such proofs do exist, but it is tricky to find them.)

A classic application of the fundamental group is to prove THE FUNDAMENTAL THEOREM OF ALGEBRA [V.13], which states that every nonconstant polynomial with complex coefficients has a complex root. (The proof is sketched in the article just cited, though the fundamental group is not explicitly mentioned there.)

The fundamental group tells us about the number of "one-dimensional holes" that a space has. A basic example is given by the circle, which has fundamental group $\mathbb{Z}$, just as $\mathbb{R}^2 \setminus \{0\}$ does, and for essentially the same reason: given a path in the circle that begins and ends at the same point, we can see how many times it goes around the circle. In the next section we shall see some more examples.

Before we think about higher-dimensional holes, we first need to discuss one of the most important topological spaces: the n-dimensional sphere. For any natural number n, this is defined to be the set of points in $\mathbb{R}^{n+1}$ at distance 1 from the origin. It is denoted S^n. Thus, the 0-sphere S^0 consists of two points, the 1-sphere S^1 is the circle, and the 2-sphere S^2 is the usual sphere, like the surface of Earth. Higher-dimensional spheres take a little bit of getting used to, but we can work with them in the same way that we can with lower-dimensional spheres. For example, we can construct the 2-sphere from a closed two-dimensional disk by identifying all the points on the boundary circle with each other. In the same way, the 3-sphere can be obtained from a solid three-dimensional ball by identifying all the points on the boundary 2-sphere. A related picture is to think of the 3-sphere as being obtained from our familiar three-dimensional space $\mathbb{R}^3$ by adding one point "at infinity."

Now let us think about the familiar sphere S^2. This has trivial fundamental group, since any loop drawn on the sphere can be shrunk to a point. However, this does not mean that the topology of S^2 is trivial. It just means that in order to detect its interesting properties we need a different invariant. And it is possible to base such an invariant on the observation that even if loops can always be shrunk, there are other maps that cannot. Indeed, the sphere itself cannot be shrunk to a point. To say this more formally, the identity map from the sphere to itself is not homotopic to a map from the sphere to just one point.

This idea leads to the notion of higher-dimensional homotopy groups of a topological space X. The rough idea is to measure the number of "n-dimensional holes" in X, for any natural number n, by considering all the continuous mappings from the n-sphere to X. We want to see whether any of these spheres wrap around a hole in X. Once again, we consider two mappings from S^n to X to be equivalent if they are homotopic. And the elements of the nth homotopy group $\pi_n(X)$ are again defined to be the homotopy classes of these mappings.

Let f be a continuous map from $[0,1]$ to X with $f(0) = f(1) = x$. If we like we can turn the interval $[0,1]$ into the circle S^1 by "identifying" the points 0 and 1: then f becomes a map from S^1 to X, with one specified point in S^1 mapping to x. In order to be able to define a group operation for mappings from a higher-dimensional S^n, we similarly fix a point s in S^n and a base point x in X and look just at maps that send s to x.

Let A and B be two continuous mappings from S^n to X with this property. The "product" mapping $A \cdot B$ from S^n to X is defined as follows. First "pinch" the equator of S^n down to a point. When $n = 1$, the equator consists of just two points and the result is a figure eight. Similarly, for general n, we end up with two copies of S^n that touch each other, one made out of the northern hemisphere and one out of the southern hemisphere of the original unpinched copy of S^n. We now use the map A to map the bottom half into X and the map B to map the top half into X, with the equator mapping to the base point x. (For both halves, the pinched equator is playing the part of the point s.)

As in the one-dimensional case, this operation makes the set $\pi_n(X)$ into a group, and this group is the nth homotopy group of the space X. One can think of it as measuring how many "n-dimensional holes" a space has.

These groups are the beginning of "algebraic" topology: starting from any topological space, we construct an algebraic object, in this case a group. If two spaces are homeomorphic, then their fundamental groups (and higher homotopy groups) must be isomorphic. This is richer than the original idea of just measuring the *number* of holes, since a group contains more information than just a number.

Any continuous function from S^n into $\mathbb{R}^m$ can be continuously shrunk to a point in a straightforward way. This shows that all the higher homotopy groups of $\mathbb{R}^m$ are also trivial, which is a precise formulation of the vague idea that $\mathbb{R}^m$ has no holes.

Under certain circumstances one can show that two different topological spaces X and Y must have the same number of holes of all types. This is clearly true if X and Y are homeomorphic, but it is also true if X and Y are equivalent in a weaker sense, known as *homotopy equivalence*. Let X and Y be topological spaces and let f_0 and f_1 be continuous maps from X to Y. A homotopy from f_0 to f_1 is defined more or less as it was for spheres: it is a continuous family of continuous maps from X to Y that starts with f_0 and ends with f_1. As then, if such a homotopy exists, we say that f_0 and f_1 are homotopic. Next, a homotopy equivalence from a space X to a space Y is a continuous map $f : X \to Y$ such that there is another continuous map $g : Y \to X$ with the property that the composition $g \circ f : X \to X$ is homotopic to the identity map on X, and $f \circ g : Y \to Y$ is homotopic to the identity map on Y. (Notice that if we replaced the word "homotopic" with "equal," we would obtain the definition of a homeomorphism.) When there is a homotopy equivalence from X to Y, we say that X and Y are *homotopy equivalent*, and also that X and Y have the same *homotopy type*.

A good example is when X is the unit circle and Y is the plane with the origin removed. We have already observed that these have the same fundamental group, and commented that it was "for essentially the same reason." Now we can be more precise. Let $f : X \to Y$ be the map that takes (x, y) to (x, y) (where the first (x, y) belongs to the circle and the second to the plane). Let $g : Y \to X$ be the map that takes (u, v) to

$$\left(\frac{u}{\sqrt{u^2+v^2}}, \frac{v}{\sqrt{u^2+v^2}}\right).$$

(Note that u^2+v^2 is never zero because the origin is not contained in Y.) Then $g \circ f$ is easily seen to equal the identity on the unit circle, so it is certainly homotopic to the identity. As for $f \circ g$, it is given by the same formula as g itself. More geometrically, it takes the points on each radial line to the point where that line intersects the unit circle. It is not hard to show that this map is homotopic to the identity on Y. (The basic idea is to "shrink the radial lines down" to the points where they intersect the circle.)

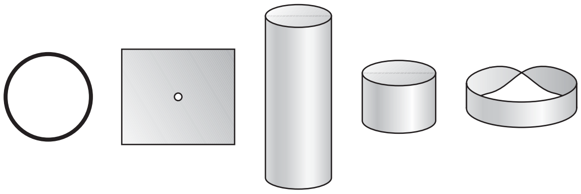

Figure 5 Some spaces that are homotopy equivalent to the circle.

Very roughly speaking, two spaces are homotopy equivalent if they have the same number of holes of all types. This is a more flexible notion of "having the same shape" than the notion of homeomorphism. For example, Euclidean spaces of different dimensions are not homeomorphic to each other, but they are all homotopy equivalent. Indeed, they are all homotopy equivalent to a point: such spaces are called *contractible*, and one thinks of them as the spaces that have no hole of any sort. The circle is not contractible, but it is homotopy equivalent to many other natural spaces: the plane $\mathbb{R}^2$ minus the origin (as we have seen), the cylinder $S^1 \times \mathbb{R}$, the compact cylinder $S^1 \times [0,1]$, and even the Möbius strip (see figure 5). Most invariants in algebraic topology (such as homotopy groups and cohomology groups) are the same for any two spaces that are homotopy equivalent. Thus, knowing that the fundamental group of the circle is isomorphic to the integers tells us that the same is true for the various homotopy equivalent spaces just mentioned. Roughly speaking, this says that all these spaces have "one basic one-dimensional hole."

3 Calculations of the Fundamental Group and Higher Homotopy Groups

To give some more feeling for the fundamental group, let us review what we already know and look at a few more examples. The fundamental group of the 2-sphere, or indeed of any higher-dimensional sphere, is trivial. The two-dimensional torus $S^1 \times S^1$ has fundamental group $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$. Thus, a loop in the torus determines two integers, which measure how many times it winds around in the meridian direction and how many in the longitudinal direction.

The fundamental group can also be non-Abelian; that is, we can have $ab \neq ba$ for some elements a and b of the fundamental group. The simplest example is a space X built out of two circles that meet at a single point (see figure 6). The fundamental group of X is the FREE GROUP [IV.10 §2] on two generators a and

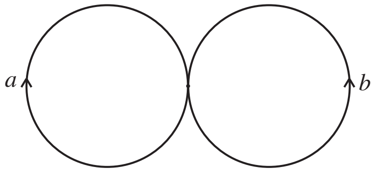

Figure 6 One-point union of two circles.

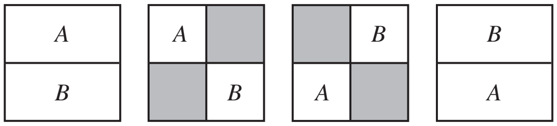

Figure 7 Proof that π_2 of any space is Abelian.

b. Roughly speaking, an element of this group is any product you can write down using the generators and their inverses, such as $abaab^{-1}a$, except that if a and a^{-1} or b and b^{-1} appear next to each other, you cancel them first. (So instead of $abb^{-1}bab^{-1}$ one would simply write $abab^{-1}$, for example.) The generators correspond to loops around each of the two circles. The free group is in a sense the most highly non-Abelian group. In particular, ab is not equal to ba, which in topological terms tells us that going around loop a and then loop b in the space X is not homotopic to the loop that goes around loop b and then loop a.

This space may seem somewhat artificial, but it is homotopy equivalent to the plane with two points removed, which appears in many contexts. More generally, the fundamental group of the plane with d points removed is the free group on d generators: this is a precise sense in which the fundamental group measures the number of holes.

In contrast with the fundamental group, the higher homotopy groups $\pi_n(X)$ are Abelian when n is at least 2. Figure 7 gives a "proof without words" in the case $n = 2$, the proof being the same for any larger n. In the figure, we view the 2-sphere as the square with its boundary identified to a point. So any elements A and B of $\pi_2(X)$ are represented by continuous maps of the square to X that map the boundary of the square to the base point x. The figure exhibits (several steps of) a homotopy from AB to BA, with the shaded regions and the boundary of the square all mapping to the base point x. The picture is reminiscent of the simplest nontrivial braid, in which one string is twisted around another; this is the beginning of a deep connection between algebraic topology and BRAID GROUPS [III.4].

The fundamental group is especially powerful in low dimensions. For example, every compact connected surface (or two-dimensional manifold) is homeomorphic to one of those on a standard list (see DIFFERENTIAL TOPOLOGY [IV.7 §2.3]), and we compute that all the manifolds on this list have different (nonisomorphic) fundamental groups. So, when you capture a closed surface in the wild, computing its fundamental group tells you exactly where it fits in the classification. Moreover, the geometric properties of the surface are closely tied to its fundamental group. The surfaces with a RIEMANNIAN METRIC [I.3 §6.10] of positive CURVATURE [III.13] (the 2-sphere and REAL PROJECTIVE PLANE [I.3 §6.7]) are exactly the surfaces with finite fundamental group; the surfaces with a metric of curvature zero (the torus and Klein bottle) are exactly the surfaces with a fundamental group that is infinite but "almost Abelian" (there is an Abelian subgroup of finite index); and the remaining surfaces, those that have a metric of negative curvature, have "highly non-Abelian" fundamental group, like the free group (see figure 8).

After more than a century of studying three-dimensional manifolds, we now know, thanks to the advances of Thurston and Perelman, that the picture is almost the same for these as it is for 2-manifolds: the fundamental group controls the geometric properties of the 3-manifold almost completely (see DIFFERENTIAL TOPOLOGY [IV.7 §2.4]). But this is completely untrue for 4-manifolds and in higher dimensions: there are many different *simply connected* manifolds, meaning manifolds with trivial fundamental group, and we need more invariants to be able to distinguish between them. (To begin with, the 4-sphere S^4 and the product $S^2 \times S^2$ are both simply connected. More generally, we can take the connected sum of any number of copies of $S^2 \times S^2$, obtained by removing 4-balls from these manifolds and identifying the boundary 3-spheres. These 4-manifolds are all simply connected, and yet no two of them are homeomorphic or even homotopy equivalent.)

An obvious way in which we might try to distinguish different spaces is to use *higher* homotopy groups, and indeed this works in simple cases. For example, π_2 of the connected sum of r copies of $S^2 \times S^2$ is isomorphic to $\mathbb{Z}^{2r}$. Also, we can show that the sphere S^n of any dimension is not contractible (although it is simply connected for $n \geqslant 2$) by computing that $\pi_n(S^n)$ is isomorphic to the integers (rather than the trivial group). Thus, each continuous map from the n-sphere to itself determines an integer, called the *degree* of the map,

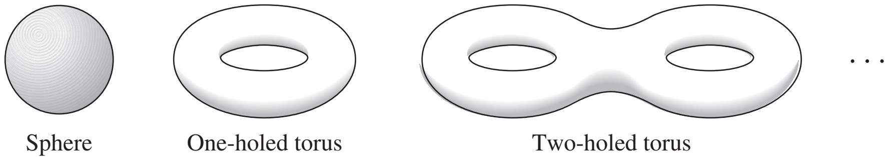

Figure 8 A sphere, a torus, and a surface of genus 2.

which generalizes the notion of winding number for maps from the circle to itself.

In general, however, the homotopy groups are not a practical way of distinguishing one space from another, because they are amazingly hard to compute. A first hint of this was Hopf's 1931 discovery that $\pi_3(S^2)$ is isomorphic to the integers: it is clear that the 2-sphere has a two-dimensional hole, as measured by $\pi_2(S^2) \cong \mathbb{Z}$, but in what sense does it have a three-dimensional hole? This does not correspond to our naive view of what such a hole should be. The problem of computing the homotopy groups of spheres turns out to be one of the hardest in all of mathematics: some of what we know is shown in table 1, but despite massive efforts the homotopy groups $\pi_i(S^2)$, for example, are known only for $i \leqslant 64$. There are tantalizing patterns in these calculations, with a number-theoretic flavor, but it seems impossible to formulate a precise guess for the homotopy groups of spheres in general. And computing the homotopy groups for spaces more complex than spheres is even more complicated.

To get an idea of the difficulties involved, let us define the so-called *Hopf map* from S^3 to S^2, which turns out to represent a nonzero element of $\pi_3(S^2)$. There are in fact several equivalent definitions. One of them is to regard a point (x_1, x_2, x_3, x_4) in S^3 as a pair of complex numbers (z_1, z_2) such that $|z_1|^2 + |z_2|^2 = 1$. This we do by setting $z_1 = x_1 + \mathrm{i}x_2$ and $z_2 = x_3 + \mathrm{i}x_4$. We then map the pair (z_1, z_2) to the complex number z_1/z_2. This may not look like a map to S^2, but it is because z_2 may be zero, so in fact the image of the map is not $\mathbb{C}$ but the *Riemann sphere* $\mathbb{C} \cup \infty$, which can be identified with S^2 in a natural way.

Another way of defining the Hopf map is to regard points (x_1, x_2, x_3, x_4) in S^3 as unit quaternions. In the article on quaternions in this volume [III.76], it is shown that each unit quaternion can be associated with a rotation of the sphere. If we fix some point s in the sphere and map each unit quaternion to the image of s under the associated rotation, then we get a map from S^3 to S^2 that is homotopic to the map defined in the previous paragraph.

The Hopf map is an important construction, and will reappear more than once later in this article.

4 Homology Groups and the Cohomology Ring

Homotopy groups, then, can be rather mysterious and very hard to calculate. Fortunately, there is a different way to measure the number of holes in a topological space: homology and cohomology groups. The definitions are more subtle than the definition of homotopy groups, but the groups turn out to be easier to compute and are for this reason much more commonly used.

Recall that elements of the nth homotopy group $\pi_n(X)$ of a topological space X are represented by continuous maps from the n-sphere to X. Let X be a manifold, for simplicity. There are two key differences between homotopy groups and homology groups. The first is that the basic objects of homology are more general than n-dimensional spheres: *every* closed oriented n-dimensional submanifold A of X determines an element of the nth homology group of X, $H_n(X)$. This might make homology groups seem much bigger than homotopy groups, but that is not the case, because of the second major difference between homotopy and homology. As with homotopy, the elements of the homology groups are not the submanifolds themselves but equivalence classes of submanifolds, but the definition of the equivalence relation for homology makes it much easier for two of these submanifolds to be equivalent than it is for two spheres to be homotopic.

We shall not give a formal definition of homology, but here are some examples that convey some of its flavor. Let X be the plane with the origin removed and let A be a circle that goes around the origin. If we continuously deform this circle, we will obtain a new curve that is homotopic to the original circle, but with homology we can do more. For instance, we can start with a continuous deformation that causes two of its points to touch and turns it into a figure eight. One half of this figure eight will have to contain the origin, but we can leave

Table 1 The first few homotopy groups of spheres.

	S^1	S^2	S^3	S^4	S^5	S^6	S^7	S^8	S^9
π_1	$\mathbb{Z}$	0	0	0	0	0	0	0	0
π_2	0	$\mathbb{Z}$	0	0	0	0	0	0	0
π_3	0	$\mathbb{Z}$	$\mathbb{Z}$	0	0	0	0	0	0
π_4	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}$	0	0	0	0	0
π_5	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}$	0	0	0	0
π_6	0	$\mathbb{Z}/4\times\mathbb{Z}/3$	$\mathbb{Z}/4\times\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}$	0	0	0
π_7	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}\times\mathbb{Z}/4\times\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}$	0	0
π_8	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/2\times\mathbb{Z}/2$	$\mathbb{Z}/8\times\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}$	0
π_9	0	$\mathbb{Z}/3$	$\mathbb{Z}/3$	$\mathbb{Z}/2\times\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/8\times\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}$
π_{10}	0	$\mathbb{Z}/3\times\mathbb{Z}/5$	$\mathbb{Z}/3\times\mathbb{Z}/5$	$\mathbb{Z}/8\times\mathbb{Z}/3\times\mathbb{Z}/3$	$\mathbb{Z}/2$	0	$\mathbb{Z}/8\times\mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$

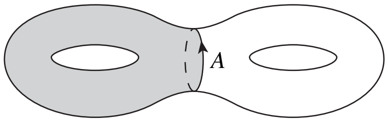

Figure 9 The circle A represents zero in the homology of the surface.

that still and slide the other part away. The result is then two closed curves, with the origin inside one and outside the other. This pair of curves, which together form a 1-manifold with two components, is equivalent to the original circle. It can be seen as a continuous deformation of a more general kind.

A second example shows how natural it is to include other manifolds in the definition of homology. This time let X be $\mathbb{R}^3$ with a circle removed, and let A be a sphere that contains the circle in its interior. Suppose that the circle is in the XY-plane and that both it and the sphere A are centered at the origin. Then we can pinch the top and bottom of A toward the origin until they just touch. If we do so, then we obtain a shape that looks like a torus, except that the hole in the middle has been shrunk to zero. But we can open up this hole with the help of a further continuous deformation and obtain a genuine torus, which is a "tube" around the original circle. From the point of view of homology, this torus is equivalent to the sphere A.

A more general rule is that if X is a manifold and B is a compact oriented $(n+1)$-dimensional submanifold of X with a boundary, then this boundary ∂B will be equivalent to zero (which is the same as saying that $[\partial B] = 0$ in $H_n(X)$): see figure 9.

The group operation is easy to define: if A and B are two disjoint submanifolds of X, giving rise to homology classes $[A]$ and $[B]$, then $[A] + [B]$ is the homology class of $[A \cup B]$. (More generally, the definition of homology allows us to add up any collection of submanifolds, whether or not they overlap.) Here are some simple examples of homology groups, which, unlike the fundamental group, are always Abelian. The homology groups of a sphere, $H_i(S^n)$, are isomorphic to the integers $\mathbb{Z}$ for $i = 0$ and for $i = n$, and 0 otherwise. This contrasts with the complicated homotopy groups of the sphere, and better reflects the naive idea that the n-sphere has one n-dimensional hole and no other holes. Note that the fundamental group of the circle, the group of integers, is the same as its first homology group. More generally, for any path-connected space, the first homology group is always the "Abelianization" of the fundamental group (which is formally defined to be its largest Abelian quotient). For example, the fundamental group of the plane with two points removed is the free group on two generators, while the first homology group is the free *Abelian* group on two generators, or $\mathbb{Z}^2$.

The homology groups of the two-dimensional torus $H_i(S^1 \times S^1)$ are isomorphic to $\mathbb{Z}$ for $i = 0$, to $\mathbb{Z}^2$ for $i = 1$, and to $\mathbb{Z}$ for $i = 2$. All of this has geometric meaning. The zeroth homology group of any space is isomorphic to $\mathbb{Z}^r$ for a space X with r connected components. So the fact that the zeroth homology group of the torus is isomorphic to $\mathbb{Z}$ means that the torus is connected. Any closed loop in the torus determines an element of the first homology group $\mathbb{Z}^2$, which measures how many times the loop winds around the meridian and longitudinal directions of the torus. And finally, the homology of the torus in dimension 2 is isomorphic to $\mathbb{Z}$ because the torus is a closed orientable manifold. That tells us that the whole torus defines an element of the second homology group of the torus, which is in fact a generator of that group. By contrast, the homotopy group

$\pi_2(S^1 \times S^1)$ is the trivial group: there are no interesting maps from the 2-sphere to the 2-torus, but homology shows that there are interesting maps from other closed 2-manifolds to the 2-torus.

As we have mentioned, calculating homology groups is much easier than calculating homotopy groups. The main reason for this is the existence of results that tell you the homology groups of a space that is built up from smaller pieces in terms of the homology groups of those pieces and their intersections. Another important property of homology groups is that they are "functorial" in the sense that a continuous map f from a space X to a space Y leads in a natural way to a homomorphism f_* from $H_i(X)$ to $H_i(Y)$ for each i: $f_*([A])$ is defined to be $[f(A)]$. In other words, $f_*([A])$ is the equivalence class of the image of A under f.

We can define the closely related idea of "cohomology" simply by a different numbering. Let X be a closed oriented n-dimensional manifold. Then we define the ith *cohomology group* $H^i(X)$ to be the homology group $H_{n-i}(X)$. Thus, one way to write down a cohomology class (an element of $H^i(X)$) is by choosing a closed oriented submanifold S of codimension i in X. (This means that the dimension of S is $n - i$.) We write $[S]$ for the corresponding cohomology class.

For more general spaces than manifolds, cohomology is not just a simple renumbering of homology. Informally, if X is a topological space, then we think of an element of $H^i(X)$ as being represented by a codimension-i subspace of X that can move around freely in X. For example, suppose that f is a continuous map from X to an i-dimensional manifold. If X is a manifold and f is sufficiently "well-behaved," then the inverse image of a "typical" point in the manifold will be an i-codimensional submanifold of X, and as we move the point about, this submanifold will vary continuously, and will do so in a way that is similar to the way that a circle became two circles and a sphere became a torus earlier. If X is a more general topological space, the map f still determines a cohomology class in $H^i(X)$, which we think of as being represented by the inverse image in X of any point in the manifold.

However, even when X is an oriented n-dimensional manifold, cohomology has distinct advantages over homology. This may seem odd, since the cohomology groups are the homology groups with different names. However, this renumbering allows us to give very useful extra algebraic structure to the cohomology groups of X: not only can we add cohomology classes, we can multiply them as well. Furthermore, we can do so in such a way that, taken together, the cohomology groups of X form a RING [III.81 §1]. (Of course, we could do this for the homology groups, but the cohomology groups form a so-called *graded* ring. In particular, if $[A] \in H^i(X)$ and $[B] \in H^j(X)$, then $[A] \cdot [B] \in H^{i+j}(X)$.)

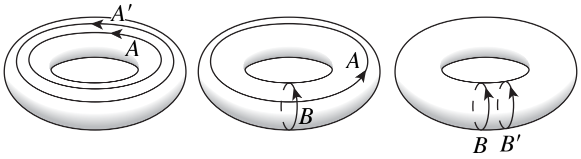

Figure 10 $A^2 = A \cdot A' = 0$, $A \cdot B = [\text{point}]$, and $B^2 = B \cdot B' = 0$.

The multiplication of cohomology classes has a rich geometric meaning, especially on manifolds: it is given by the *intersection* of two submanifolds. This generalizes our discussion of intersection numbers in section 1: there we considered zero-dimensional intersections of submanifolds, whereas we are now considering (cohomology classes of) higher-dimensional intersections. To be precise, let S and T be closed oriented submanifolds of X, of codimension i and j, respectively. By moving S slightly (which does not change its class in $H^i(X)$) we can assume that S and T intersect transversely, which implies that the intersection of S and T is a smooth submanifold of codimension $i + j$ in X. Then the product of the cohomology classes $[S]$ and $[T]$ is simply the cohomology class of the intersection $S \cap T$ in $H^{i+j}(X)$. (In addition, the submanifold $S \cap T$ inherits an orientation from S, T, and X: this is needed to define the associated cohomology class.)

As a result, to compute the cohomology ring of a manifold, it is enough to specify a basis for the cohomology groups (which, as we have already discussed, are relatively easy to determine) using some submanifolds and to see how these submanifolds intersect. For example, we can compute the cohomology ring of the 2-torus as shown in figure 10. For another example, it is not hard to show that the cohomology of the COMPLEX PROJECTIVE PLANE [III.72] $\mathbb{CP}^2$ has a basis given by three basic submanifolds: a point, which belongs to $H^4(\mathbb{CP}^2)$ because it is a submanifold of codimension 4; a complex projective line $\mathbb{CP}^1 = S^2$, which belongs to $H^2(\mathbb{CP}^2)$; and the whole manifold $\mathbb{CP}^2$, which is in $H^0(\mathbb{CP}^2)$ and represents the identity element 1 of the cohomology ring. The product in the cohomology ring is described by saying that $[\mathbb{CP}^1][\mathbb{CP}^1] = [\text{point}]$, because any two distinct lines $\mathbb{CP}^1$ in the plane meet transversely in a single point.

This calculation of the cohomology ring of the complex projective plane, although very simple, has several strong consequences. First of all, it implies Bézout's theorem on intersections of complex algebraic curves (see ALGEBRAIC GEOMETRY [IV.4 §6]). An algebraic curve of degree d in $\mathbb{CP}^2$ represents d times the class of a line $\mathbb{CP}^1$ in $H^2(\mathbb{CP}^2)$. Therefore, if two algebraic curves D and E of degrees d and e meet transversely, then the cohomology class $[D \cap E]$ equals

$$[D] \cdot [E] = (d[\mathbb{CP}^1])(e[\mathbb{CP}^1]) = de[\text{point}].$$

For complex submanifolds of a complex manifold, intersection numbers are always $+1$, not -1, and so this means that D and E meet in exactly de points.

We can also use the computation of the cohomology ring of $\mathbb{CP}^2$ to prove something about the homotopy groups of spheres. It turns out that $\mathbb{CP}^2$ can be constructed as the union of the 2-sphere and the closed four-dimensional ball, with each point of the boundary S^3 of the ball identified with a point in S^2 by the Hopf map, which was defined in the previous section.

A constant map from one space to another, or a map homotopic to a constant map, gives rise to the zero homomorphism between the homology groups H_i, at least when $i > 0$. The Hopf map $f : S^3 \to S^2$ also induces the zero homomorphism because the nonzero homology groups of S^3 and S^2 are in different dimensions. Nonetheless, we will show that f is not homotopic to the constant map. If it were, then the space $\mathbb{CP}^2$ obtained by attaching a 4-ball to the 2-sphere using the map f would be homotopy equivalent to the space obtained by attaching a 4-ball to the 2-sphere using a constant map. The latter space Y is the union of S^2 and S^4 identified at one point. But in fact Y is not homotopy equivalent to the complex projective plane, because their cohomology rings are not isomorphic. In particular, the product of any element of $H^2(Y)$ with itself is zero, unlike what happens in $\mathbb{CP}^2$ where $[\mathbb{CP}^1][\mathbb{CP}^1] = [\text{point}]$. Therefore f is nonzero in $\pi_3(S^2)$. A more careful version of this argument shows that $\pi_3(S^2)$ is isomorphic to the integers, and the Hopf map $f : S^3 \to S^2$ is a generator of this group.

This argument shows some of the rich relations between all the basic concepts of algebraic topology: homotopy groups, cohomology rings, manifolds, and so on. To conclude, here is a way to visualize the nontriviality of the Hopf map $f : S^3 \to S^2$. Look at the subset of S^3 that maps to any given point of the 2-sphere. These inverse images are all circles in the 3-sphere. To draw them, we can use the fact that S^3 minus a point is homeomorphic to $\mathbb{R}^3$; so these inverse images form a family of disjoint circles that fills up three-dimensional space, with one circle being drawn as a line (the circle through the point we removed from S^3). The striking feature of this picture is that any two of this huge family of circles have linking number 1 with each other: there is no way to pull any two of them apart (see figure 11).

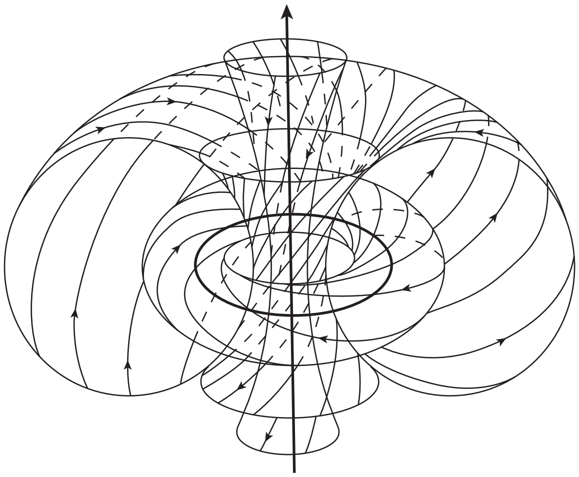

Figure 11 Fibers of the Hopf map.

5 Vector Bundles and Characteristic Classes

We now introduce another major topological idea: fiber bundles. If E and B are topological spaces, x is a point in B, and $p : E \to B$ is a continuous map, then the *fiber* of p *over* x is the subspace of E that maps to x. We say that p is a *fiber bundle*, with fiber F, if every fiber of p is homeomorphic to the same space F. We call B the *base space* and E the *total space*. For example, any product space $B \times F$ is a fiber bundle over B, called the trivial F-bundle over B. (The continuous map in this case is the map that takes (x, y) to x.) But there are many nontrivial fiber bundles. For example, the Möbius strip is a fiber bundle over the circle with fiber a closed interval. This example helps to explain the old name "twisted product" for fiber bundles. Another example: the Hopf map makes the 3-sphere the total space of a circle bundle over the 2-sphere.

Fiber bundles are a fundamental way to build up complicated spaces from simple pieces. We will focus on the most important special case: vector bundles. A *vector bundle* over a space B is a fiber bundle $p : E \to B$ whose fibers are all real vector spaces of some dimension n.

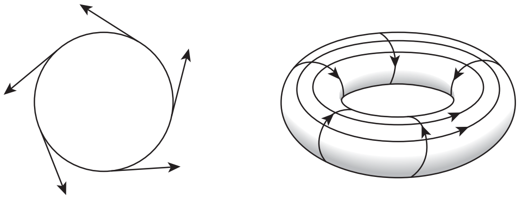

Figure 12 Trivializations of the tangent bundle for the circle and the torus.

This dimension is called the *rank* of the vector bundle. A *line bundle* means a vector bundle of rank 1; for example, we can view the Möbius strip (not including its boundary) as a line bundle over the circle S^1. It is a *nontrivial* line bundle; that is, it is not isomorphic to the trivial line bundle $S^1 \times \mathbb{R}$. (There are many ways of constructing it: one is to take the strip $\{(x, y) : 0 \leqslant x \leqslant 1\}$ and identify each point $(0, y)$ with the point $(1, -y)$. The base space of this line bundle is the set of all points $(x, 0)$, which is a circle since $(0, 0)$ and $(1, 0)$ have been identified.)

If M is a smooth manifold of dimension n, its *tangent bundle* $TM \to M$ is a vector bundle of rank n. We can easily define this bundle by considering M as a submanifold of some Euclidean space $\mathbb{R}^N$. (Every smooth manifold can be embedded into Euclidean space.) Then TM is the subspace of $M \times \mathbb{R}^N$ of pairs (x, v) such that the vector v is tangent to M at the point x; the map $TM \to M$ sends a pair (x, v) to the point x. The fiber over x then has the form of the set of all pairs (x, v) with v belonging to an affine subspace of $\mathbb{R}^N$ of dimension equal to that of M. For any fiber bundle, a *section* means a continuous map from the base space B to the total space E that maps each point x in B to some point in the fiber over x. A section of the tangent bundle of a manifold is called a *vector field*. We can draw a vector field on a given manifold by putting an arrow (possibly of zero length) at every point of the manifold.

In order to classify smooth manifolds, it is important to study their tangent bundles, and in particular to see whether they are trivial or not. Some manifolds, like the circle S^1 and the torus $S^1 \times S^1$, do have trivial tangent bundle. The tangent bundle of an n-manifold M is trivial if and only if we can find n vector fields that are linearly independent at every point of M. So we can prove that the tangent bundle is trivial just by writing down such vector fields; see figure 12 for the circle or the torus. But how can we show that the tangent bundle of a given manifold is nontrivial?

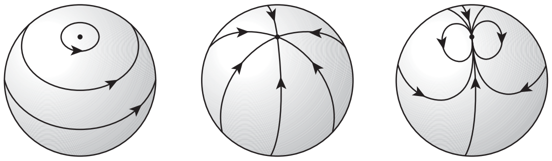

Figure 13 The hairy ball theorem.

One way is to use intersection numbers. Let M be a closed oriented n-manifold. We can identify M with the image of the "zero-section" inside the tangent bundle TM, the section that assigns to every point of M the zero vector at that point. Since the dimension of TM is precisely double that of M, the discussion of intersection numbers in section 1 gives a well-defined integer $M^2 = M \cdot M$, the self-intersection number of M inside TM; this is called the *Euler characteristic* $\chi(M)$. By the definition of intersection numbers, for any vector field v on M that meets the zero-section transversely, the Euler characteristic of M is equal to the number of zeros of v, counted with signs.

As a result, if the Euler characteristic of M is not zero, then every vector field on M must meet the zero-section; in other words, every vector field on M must equal zero somewhere. The simplest example occurs when M is the 2-sphere S^2. We can easily write down a vector field (for example, the one pointing toward the east along circles of latitude, which vanishes at the north and south poles) whose intersection number with the zero-section is 2. Therefore the 2-sphere has Euler characteristic 2, and so every vector field on the 2-sphere must vanish somewhere. This is a famous theorem of topology known as the "hairy ball theorem": it is impossible to comb the hair on a coconut (see figure 13).

This is the beginning of the theory of *characteristic classes*, which measure how nontrivial a given vector bundle is. There is no need to restrict ourselves to the tangent bundle of a manifold. For any oriented vector bundle E of rank n on a topological space X, we can define a cohomology class $\chi(E)$ in $H^n(X)$, the *Euler class*, which vanishes if the bundle is trivial. Intuitively, the Euler class of E is the cohomology class represented by the zero set of a general section of E, which (for example, if X is a manifold) should be a codimension-n submanifold of X, since X has codimension n in E. If X is a closed oriented n-manifold, then the Euler class of the tangent bundle in $H^n(X) = \mathbb{Z}$ is the Euler characteristic of X.

One of the inspirations for the theory of characteristic classes was the Gauss–Bonnet theorem, generalized to all dimensions in the 1940s. The theorem expresses the Euler characteristic of a closed manifold with a Riemannian metric as the integral over the manifold of a certain curvature function. More broadly, a central goal of differential geometry is to understand how the geometric properties of a Riemannian manifold such as its curvature are related to the topology of the manifold.

The characteristic classes for *complex* vector bundles (that is, bundles where the fibers are complex vector spaces) turn out to be particularly convenient: indeed, real vector bundles are often studied by constructing the associated complex vector bundle. If E is a complex vector bundle of rank n over a topological space X, the *Chern classes* of E are a sequence $c_1(E), \dots, c_n(E)$ of cohomology classes on X, with $c_i(E)$ belonging to $H^{2i}(X)$, which all vanish if the bundle is trivial. The top Chern class, $c_n(E)$, is simply the Euler class of E: thus, it is the first obstruction to finding a section of E that is everywhere nonzero. The more general Chern classes have a similar interpretation. For any $1 \leqslant j \leqslant n$, choose j general sections of E. The subset of X over which these sections become linearly dependent will have codimension $2(n+1-j)$ (assuming, for example, that X is a manifold). The Chern class $c_{n+1-j}(E)$ is precisely the cohomology class of this subset. Thus the Chern classes measure in a natural way the failure of a given complex vector bundle to be trivial. The *Pontryagin classes* of a real vector bundle are defined to be the Chern classes of the associated complex vector bundle.

A triumph of differential topology is Sullivan's 1977 theorem that there are only finitely many smooth closed simply connected manifolds of dimension at least 5 with any given homotopy type and given Pontryagin classes of the tangent bundle. This statement fails badly in dimension 4, as Donaldson discovered in the 1980s (see DIFFERENTIAL TOPOLOGY [IV.7 §2.5]).

6 *K*-Theory and Generalized Cohomology Theories

The effectiveness of vector bundles in geometry led to a new way of measuring the "holes" in a topological space X: looking at how many different vector bundles over X there are. This idea gives a simple way to define a cohomology-like ring associated to any space, known as K-theory (after the German word "Klasse," since the theory involves equivalence classes of vector bundles). It turns out that K-theory gives a very useful new angle by which to look at topological spaces. Some problems that could be solved only with enormous effort using ordinary cohomology became easy with K-theory. The idea was created in algebraic geometry by Grothendieck in the 1950s and then brought into topology by Atiyah and Hirzebruch in the 1960s.

The definition of K-theory can be given in a few lines. For a topological space X, we define an Abelian group $K^0(X)$, the *K-theory* of X, whose elements can be written as formal differences $[E]-[F]$, where E and F are any two complex vector bundles over X. The only relations we impose in this group are that $[E \oplus F] = [E]+[F]$ for any two vector bundles E and F over X. Here $E \oplus F$ denotes the *direct sum* of the two bundles; if E_x and F_x denote the fibers at a given point x in X, the fiber of $E \oplus F$ at x is simply $E_x \times F_x$.

This simple definition leads to a rich theory. First of all, the Abelian group $K^0(X)$ is in fact a ring: we multiply two vector bundles on X by forming the TENSOR PRODUCT [III.89]. In this respect, K-theory behaves like ordinary cohomology. The analogy suggests that the group $K^0(X)$ should form part of a whole sequence of Abelian groups $K^i(X)$, for integers i, and indeed these groups can be defined. In particular, $K^{-i}(X)$ can be defined as the subgroup of those elements of $K^0(S^i \times X)$ whose restriction to $K^0(\text{point} \times X)$ is zero.

Then a miracle occurs: the groups $K^i(X)$ turn out to be *periodic* of order 2: $K^i(X) = K^{i+2}(X)$ for all integers i. This is a famous phenomenon known as *Bott periodicity*. So there are really only two different K-groups attached to any topological space: $K^0(X)$ and $K^1(X)$.

This may suggest that K-theory contains less information than ordinary cohomology, but that is not so. Neither K-theory nor ordinary cohomology determines the other, although there are strong relations between them. Each brings different aspects of the shape of a space to the fore. Ordinary cohomology, with its numbering, shows fairly directly the way a space is built up from pieces of different dimensions. K-theory, having only two different groups, looks cruder at first (and is often easier to compute as a result). But geometric problems involving vector bundles often involve information that is subtle and hard to extract from ordinary cohomology, whereas this information is brought to the surface by K-theory.

The basic relation between K-theory and ordinary cohomology is that the group $K^0(X)$ constructed from the vector bundles on X "knows" something about all the even-dimensional cohomology groups of X. To be precise, the rank of the Abelian group $K^0(X)$ is the sum of the ranks of all the even-dimensional cohomology groups $H^{2i}(X)$. This connection comes from associating with a given vector bundle on X its Chern classes. The odd K-group $K^1(X)$ is related in the same way to the odd-dimensional ordinary cohomology.

As we have already hinted, the precise group $K^0(X)$, as opposed to just its rank, is better adapted to some geometric problems than ordinary cohomology. This phenomenon shows the power of looking at geometric problems in terms of vector bundles, and thus ultimately in terms of linear algebra. Among the classic applications of K-theory is the proof, by Bott, Kervaire, and Milnor, that the 0-sphere, the 1-sphere, the 3-sphere, and the 7-sphere are the only spheres whose tangent bundles are trivial. This has a deep algebraic consequence, in the spirit of the fundamental theorem of algebra: the only dimensions in which there can be a real division algebra (not assumed to be commutative or even associative) are 1, 2, 4, and 8. There are indeed division algebras of all four types: the real numbers, complex numbers, quaternions, and octonions (see QUATERNIONS, OCTONIONS, AND NORMED DIVISION ALGEBRAS [III.76]).

Let us see why the existence of a real division algebra of dimension n implies that the $(n-1)$-sphere has trivial tangent bundle. In fact, let us merely assume that we have a finite-dimensional real vector space V with a bilinear map $V \times V \to V$, which we call the "product," such that if x and y are vectors in V with $xy = 0$, then either $x = 0$ or $y = 0$. For convenience, let us also assume that there is an identity element 1 in V, so $1 \cdot x = x \cdot 1 = x$ for all $x \in V$; one can, however, do without this assumption. If V has dimension n, then we can identify V with $\mathbb{R}^n$. Then, for each point x in the sphere S^{n-1}, left multiplication by x gives a linear isomorphism from $\mathbb{R}^n$ to itself. By scaling the output to have length 1, left multiplication by x gives a diffeomorphism from S^{n-1} to itself which maps the point 1 (scaled to have length 1) to x. Taking the derivative of this diffeomorphism at the point 1 gives a linear isomorphism from the tangent space of the sphere at the point 1 to the tangent space at x. Since the point x on the sphere is arbitrary, a choice of basis for the tangent space of the sphere at the point 1 determines a trivialization of the whole tangent bundle of the $(n-1)$-sphere.

Among other applications, K-theory provides the best "explanation" for the low-dimensional homotopy groups of spheres, and in particular for the number-theoretic patterns that are seen there. Notably, denominators of Bernoulli numbers appear among those groups (such as $\pi_{n+3}(S^n) \cong \mathbb{Z}/24$ for n at least 5), and this pattern was explained using K-theory by Milnor, Kervaire, and Adams.

THE ATIYAH–SINGER INDEX THEOREM [V.2] provides a deep analysis of linear differential equations on closed manifolds using K-theory. The theorem has made K-theory important for gauge theories and string theories in physics. K-theory can also be defined for noncommutative rings, and is in fact the central concept in "noncommutative geometry" (see OPERATOR ALGEBRAS [IV.15 §5]).

The success of K-theory led to a search for other "generalized cohomology theories." There is one other theory that stands out for its power: *complex cobordism*. The definition is very geometric: the complex cobordism groups of a manifold M are generated by mappings of manifolds (with a complex structure on the tangent bundle) into M. The relations say that any manifold counts as zero if it is the boundary of some other manifold. For example, the union of two circles would count as zero if you could find a cylinder whose ends were those circles.

It turns out that complex cobordism is much richer than either K-theory or ordinary cohomology. It sees far into the structure of a topological space, but at the cost of being difficult to compute. Over the past thirty years, a whole series of cohomology theories, such as elliptic cohomology and Morava K-theories, have been constructed as "simplifications" of complex cobordism: there is a constant tension in topology between invariants that carry a lot of information and invariants that are easy to compute. In one direction, complex cobordism and its variants provide the most powerful tool for the computation and understanding of the homotopy groups of spheres. Beyond the range where Bernoulli numbers appear, we see deeper number theory such as MODULAR FORMS [III.59]. In another direction, the geometric definition of complex cobordism makes it useful in algebraic geometry.

7 Conclusion

The line of thought introduced by pioneering topologists like RIEMANN [VI.49] is simple but powerful. Try to translate any problem, even a purely algebraic one, into geometric terms. Then ignore the details of the geometry and study the underlying shape or topology of the problem. Finally, go back to the original problem and see how much has been gained. The fundamental topological ideas such as cohomology are used throughout mathematics, from number theory to string theory.

Further Reading

From the definition of topological spaces to the fundamental group and a little beyond, I like M. A. Armstrong's *Basic Topology* (Springer, New York, 1983). The current standard graduate textbook is A. Hatcher's *Algebraic Topology* (Cambridge University Press, Cambridge, 2002). Two of the great topologists, Bott and Milnor, are also brilliant writers. Every young topologist should read R. Bott and L. Tu's *Differential Forms in Algebraic Topology* (Springer, New York, 1982), J. Milnor's *Morse Theory* (Princeton University Press, Princeton, NJ, 1963), and J. Milnor and J. Stasheff's *Characteristic Classes* (Princeton University Press, Princeton, NJ, 1974).

IV.7 Differential Topology
C. H. Taubes

1 Smooth Manifolds

This article is about classifying certain objects called smooth manifolds, so I need to start by telling you what they are. A good example to keep in mind is the surface of a smooth ball. If you look at a small portion of it from very close up, then it looks like a portion of a flat plane, but of course it differs in a radical way from a flat plane on larger distance scales. This is a general phenomenon: a smooth manifold can be very convoluted, but must be quite regular in close-up. This "local regularity" is the condition that each point in a manifold belongs to a neighborhood that looks like a portion of standard Euclidean space in some dimension. If the dimension in question is d for every point of the manifold, then the manifold itself is said to have dimension d. A schematic of this is shown in figure 1.

What does it mean to say that a neighborhood "looks like a portion of standard Euclidean space"? It means that there is a "nice" one-to-one map ϕ from the neighborhood into $\mathbb{R}^d$ (with its usual notion of distance). One can think of ϕ as "identifying" points in the neighborhood with points in $\mathbb{R}^d$: that is, x is identified with $\phi(x)$. If we do this, then the function ϕ is called a *coordinate chart* of the neighborhood, and any chosen basis for the linear functions on the Euclidean space is called a *coordinate system*. The reason for this is that ϕ allows us to use the coordinates in $\mathbb{R}^d$ to label points in the neighborhood: if x belongs to the neighborhood, then one can label it with the coordinates of $\phi(x)$. For example, Europe is part of the surface of a sphere. A typical map of Europe identifies each point in Europe with a point in flat, two-dimensional Euclidean space, that is, a square grid labeled with latitude and longitude. These two numbers give us a coordinate system for the map, which can also be transferred to a coordinate system for Europe itself.

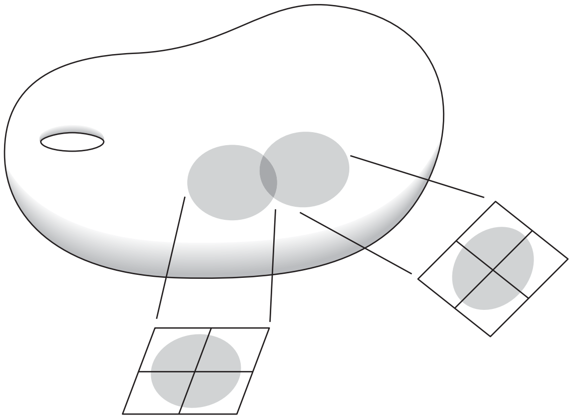

Figure 1 Small portions of a manifold resemble regions in a Euclidean space.

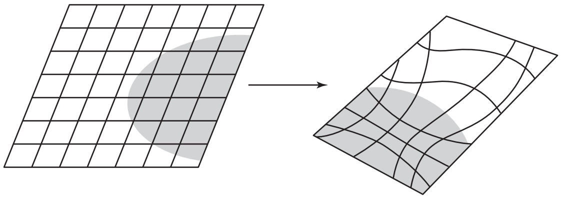

Figure 2 A transition function from a rectangular grid to a distorted rectangular grid.

Now, here is a straightforward but central observation. Suppose that M and N are two neighborhoods that intersect, and suppose that functions $\phi : M \to \mathbb{R}^d$ and $\psi : N \to \mathbb{R}^d$ are used to give them each a coordinate chart. Then the intersection $M \cap N$ is given *two* coordinate charts, and this gives us an identification between the open regions $\phi(M \cap N)$ and $\psi(M \cap N)$ of $\mathbb{R}^d$: given a point x in the first region, the corresponding point in the second is $\psi(\phi^{-1}(x))$. This composition of maps is called a *transition function*, and it tells you how the coordinates from one of the charts on the intersecting region relate to those of the other. The transition function is a HOMEOMORPHISM [III.90] between the regions $\phi(M \cap N)$ and $\psi(M \cap N)$.

Suppose that you take a rectangular grid in the first Euclidean region and use the transition function $\psi\phi^{-1}$ to map it to the second one. It is possible that the image will again be a rectangular grid, but in general it will be somewhat distorted. An illustration is given in figure 2.

The proper term for a space whose points are surrounded by regions that can be identified with parts of Euclidean space is a *topological manifold*. The word "topological" is used in order to indicate that there are no constraints on the coordinate-chart transition functions apart from the basic one that they should be continuous. However, some continuous functions are quite unpleasant, so one typically introduces extra constraints in order to limit the distorting effect that the transition functions can have on a rectangular coordinate grid.

Of prime interest here is the case where the transition functions are required to be differentiable to all orders. If a manifold has a collection of charts for which all the transition functions are infinitely differentiable, then it is said to have a *smooth structure*, and it is called a *smooth manifold*. Smooth manifolds are especially interesting because they are the natural arena for calculus. Roughly speaking, they are the most general context in which the notion of differentiation to any order makes intrinsic sense.

A function f, defined on a manifold, is said to be *differentiable* if, given any of its coordinate charts $\phi : N \to \mathbb{R}^d$, the function $g(y) = f(\phi^{-1}(y))$ (which is defined on a region of $\mathbb{R}^d$) is DIFFERENTIABLE [I.3 §5.3]. Calculus is impossible on a manifold if it does not admit charts with differentiable transition functions, because a function that might appear differentiable in one chart will not, in general, be differentiable when viewed from a neighboring chart.

Here is a one-dimensional example to illustrate this point. Consider the following two coordinate charts of a neighborhood of the origin in the real line. The first is the obvious chart that simply represents a real number x by itself. (Formally speaking, one is taking the function ϕ to be defined by the simple formula $\phi(x) = x$.) The second represents x by the point $x^{1/3}$. (Here the cube root of a negative number x is defined to be minus the cube root of $-x$.) What is the transition function between these two charts? Well, if t is a point in the region of $\mathbb{R}$ used for the first chart, then $\phi^{-1}(t) = t$, so $\psi(\phi^{-1}(t)) = \psi(t) = t^{1/3}$. This is a continuous function of t but it is not differentiable at the origin.

Now consider the simplest possible function defined on the region used for the second chart, the function $h(s) = s$, and let us work out the corresponding function f on the manifold itself. The value of f at x should be the value of h at the point s corresponding to x. This point is $\psi(x) = x^{1/3}$, so $f(x) = h(x^{1/3}) = x^{1/3}$. Finally, since the point x in the manifold corresponds to the point $t = \phi(x) = x$ in the first region, the corresponding function on the first region is $g(t) = t^{1/3}$. (This is the same function as f only because ϕ happens to be the very special map that takes each number to itself.) Thus, the eminently differentiable function h on one coordinate chart translates into the continuous but not differentiable function g on the other.

Suppose one is given a topological manifold M with two sets of charts, both of which have infinitely differentiable transition functions. Then each set of charts gives us a smooth structure on the manifold. Of great importance is the fact that these two smooth structures can be fundamentally different.

To see what this means, let us call the sets of charts K and L. Given a function f, let us call it *K-differentiable* if it is differentiable from the viewpoint of K, and *L-differentiable* if it is differentiable from the viewpoint of L. It may easily happen that a function is K-differentiable without being L-differentiable or vice versa. However, we can say that K and L *give the same smooth structure* on M when there is a map, F, from M to itself with the following three properties. First, F is invertible and both F and F^{-1} are continuous. Second, the composition of F with any function that is K-differentiable is L-differentiable. Third, the composition of the inverse of F^{-1} with any function that is L-differentiable is K-differentiable. Loosely speaking, F turns the K-differentiable functions into L-differentiable ones and F^{-1} turns them back again. If no such function F exists, then the smooth structures given by K and L are considered to be genuinely different.

To see how this plays out, let us look at the one-dimensional example again. As noted previously, the functions that you deem to be differentiable if you use the ϕ-chart are not the same as those you deem to be differentiable if you use the ψ-chart. For example, the function $x \mapsto x^{1/3}$ is not ϕ-differentiable but it is ψ-differentiable. Even so, the ϕ-differentiable and ψ-differentiable sets of functions define the *same* smooth structure for the line, since any ψ-differentiable function becomes ϕ-differentiable once you compose it with the self-map $F : t \mapsto t^3$.

It is very far from obvious that any manifold can have more than one smooth structure, but this turns out to be the case. There are also manifolds that are entirely lacking in smooth structures. These two facts lead directly to the central concern of this essay, the long-sought quest for the two holy grails of differential topology.

- A list of all smooth structures on any given topological manifold.
- An algorithm to identify any given smooth structure on any given topological manifold with the corresponding structure from the list.

2 What Is Known about Manifolds?

Much has been accomplished as of the writing of this article with respect to the two points listed above. This said, the task for this part of the article is to summarize the state of affairs at the beginning of the twenty-first century. Various examples of manifolds are described along the way.

The story here requires a brief, preliminary digression to set the stage. If you have two manifolds and you set them side by side without their touching, then technically speaking they can be regarded as a single manifold that happens to have two components. In such a case, one can study the components individually. Therefore, in this article I shall talk exclusively about *connected* manifolds: that is, manifolds with just one component. In a connected manifold, one can get from any point to any other point without ever leaving the manifold.

A second technical point is that it is useful to distinguish between manifolds such as the sphere, which are bounded in extent, and manifolds such as the plane, which go off to infinity. More precisely, I am talking about the distinction between COMPACT [III.9] and noncompact manifolds: a compact manifold can be thought of as one that can be expressed as a closed bounded subset of $\mathbb{R}^n$ for some n. The discussion that follows will be almost entirely about compact manifolds. As some of the examples below will demonstrate, the story for compact manifolds is less convoluted than the analogous story for noncompact ones. For simplicity I shall often use the word "manifold" to mean "compact manifold"; it will be clear from the context if noncompact manifolds are also being discussed.

2.1 Dimension 0

There is only one dimension-0 manifold. It is a single point. The period at the end of this sentence looks, from afar, like a connected, dimension-0 manifold. Note that the distinction between topological and smooth is irrelevant here.

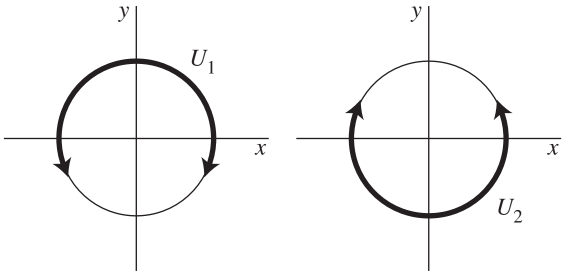

Figure 3 Two charts that cover the circle.

2.2 Dimension 1

There is only one compact, connected, one-dimensional topological manifold, namely the circle. Moreover, the circle has just one smooth structure. Here is one way to represent this structure. Take as a representative circle the unit circle in the xy-plane, that is, the set of all points (x, y) with $x^2 + y^2 = 1$. This can be covered by two overlapping intervals, each of which covers slightly more than half of the circle. The intervals U_1 and U_2 are drawn in figure 3. Each interval constitutes a coordinate chart. The one on the left, U_1, can be parametrized in a continuous fashion by taking the angle of a given point as measured counterclockwise from the positive x-axis. For example, the point $(1, 0)$ has angle 0, and the point $(-1, 0)$ has angle π. In order to parametrize U_2 by angle, you will have to start with angle π at the negative x-axis. If you move around U_2, varying this angle continuously, then when you reach the point $(1, 0)$ you will have parametrized it as a point in U_2 using the angle 2π.

As you can see, the arcs U_1 and U_2 intersect in two separated, smaller arcs; these are labeled V_1 and V_2 in figure 4. The transition function on V_1 is the identity map, since the U_1 angle of any given point in V_1 is the same as its U_2 angle. By contrast, the U_2 angle of a point in V_2 is obtained from the U_1 angle by adding 2π. Thus, the transition function on V_2 is not the identity map but the map that adds 2π to the coordinate function.

This one-dimensional example brings up a number of important issues, all related to a particularly troubling question. To state it, consider first that there are lots of closed loops in the plane that can be taken as model circles. Indeed, the word "lots" considerably understates the situation. Moreover, why should we restrict our

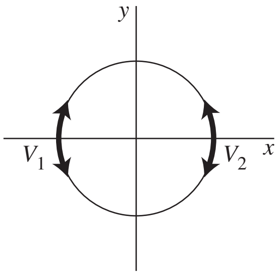

Figure 4 The intersection of the arcs U_1 and U_2.

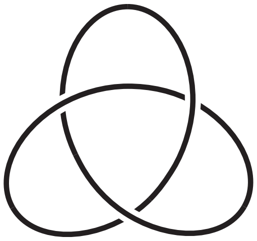

Figure 5 A knotted loop in 3-space.

attention to circles in the plane? There are closed loops galore in 3-space too: see figure 5, for example. For that matter, any manifold of dimension greater than 1 has smooth loops. Earlier, it was asserted that there is just one smooth, compact, connected, one-dimensional manifold, so all of these loops must be considered the "same." Why is this?

Here is the answer. We often think of a manifold as it might appear were it sitting in some larger space. For example, we might imagine a circle sitting in the plane, or sitting knotted in three-dimensional Euclidean space. However, the notion of "smooth manifold" introduced above is an *intrinsic* one, in the sense that it does not depend on how the manifold is placed inside a higher-dimensional space. Indeed, it is not even necessary for there to be a higher-dimensional space at all. In the case of the circle, this can be said in the following way. The circle can be placed as a loop in the plane, or as a knot in 3-space, or whatever. Each view of the circle in a higher-dimensional Euclidean space defines a collection of functions that are considered differentiable: one just takes the differentiable functions of the coordinates of the big Euclidean space and restricts them to the circle. As it turns out, any one such collection defines the same smooth structure on the circle as any other. Thus, the smooth structures that are provided by these different views of a circle are all the same, even though there are many interesting ways of placing a circle in a given higher-dimensional space. (In fact, the classification of knots in 3-space is a fascinating, vibrant topic in its own right: see KNOT POLYNOMIALS [III.44].)

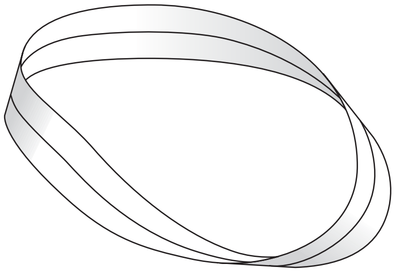

Figure 6 A Möbius strip has just one side.

How is it proved that there is only one smooth structure for the circle? For that matter, how is it proved that there is but a single compact topological manifold in dimension 1? Since this article is not meant to provide proofs, these questions are left as serious exercises with the following advice. Think hard about the definitions and, for the smooth-manifold question, use some calculus.

2.3 Dimension 2

The story for two-dimensional, connected, compact manifolds is much richer than that for dimension 1. In the first place, there is a basic dichotomy between two kinds of manifold: orientable and nonorientable. Roughly speaking, this is the distinction between manifolds that have two sides and those that have just one. To give a more formal definition, a two-dimensional manifold is called *orientable* if every loop in the manifold that does not cross itself or have any kinks has two distinct sides. This is to say that there is no path from one side of the loop to the other that avoids the loop yet remains very close to it. The Möbius strip (see figure 6) is not orientable because there are paths from one side of the central loop to the other that do not cross the central loop yet remain very close to it. The orientable, compact, connected, topological, two-dimensional manifolds are in one-to-one correspondence with a collection of fundamental foods: the apple, the doughnut, the two-holed pretzel, the three-holed pretzel, the four-holed pretzel, and so on (see figure 7). Technically, they are classified by an integer, called the *genus*. This is 0 for the sphere, 1 for the torus, 2 for the two-holed torus, etc. The genus counts the number of holes that appear in a given example from

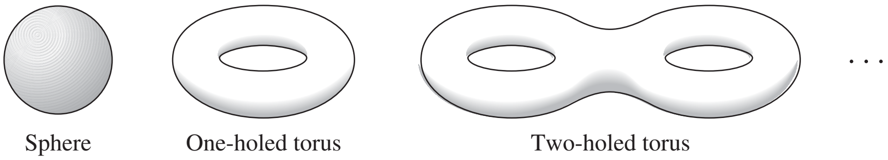

Figure 7 The orientable manifolds of dimension 2.

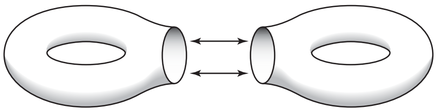

Figure 8 Cutting and gluing.

figure 7. To say that this classifies them is to say that two such manifolds are the same if and only if they have the same genus. This is a theorem due to POINCARÉ [VI.61].

As it turns out, every topological two-dimensional manifold has exactly one smooth structure, so the list in figure 7 is the same as the list of the *smooth* orientable two-dimensional manifolds. Here one should keep in mind that the notion of a smooth manifold is intrinsic, and therefore independent of how the manifold is represented as a surface in 3-space, or in any other space. For example, the surfaces of an orange, a banana, and a watermelon each represent embedded images of the two-dimensional sphere, the leftmost example in figure 7.

The shapes illustrated in figure 7 suggest an idea that plays a key role when it comes to classifying manifolds of higher dimensions. Notice that the two-holed torus can be viewed as the result of taking two one-holed tori, cutting disks out of both, gluing the results together across their boundary circles, and then smoothing the corners. This operation is depicted in figure 8. This sort of cutting and gluing operation is an example of what is called a *surgery*. The analogous surgery can also be done with a one-holed torus and a two-holed torus to obtain a three-holed torus. And so on. Thus, all of the oriented two-dimensional manifolds can be built using standard surgeries on copies of just two fundamental building blocks: the one-holed torus and the sphere. Here is a nice exercise to test your understanding of this process. Suppose that you perform a surgery, as in figure 8, on a sphere and another manifold M. Prove that the resulting manifold is the same, with regard to its topological and smooth structure, as M.

As it turns out, all of the nonorientable two-dimensional manifolds can be built using a version of surgery that first cuts a disk out of an orientable two-dimensional manifold and then glues on a Möbius strip. To be more precise, note that the Möbius strip has a circle as its boundary. Cut a disk out of any given orientable, two-dimensional manifold and the result also has a circular boundary. Glue the latter circular boundary to the Möbius strip boundary, smooth the corners, and the result is a smooth manifold that is nonorientable. Every nonorientable, topological (and thus every nonorientable, smooth), two-dimensional manifold is obtained in this way. Moreover, the manifold you get depends only on the number of holes (the genus) of the orientable manifold that is used.

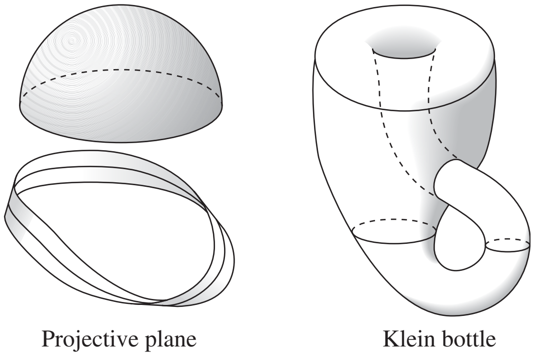

Figure 9 Two nonorientable surfaces. To form the projective plane, one identifies the boundary of the Möbius strip with the boundary of the hemisphere.

The manifold obtained from the surgery of a Möbius strip with a sphere is called the *projective plane*. The one that uses the Möbius strip and the torus is called the *Klein bottle*. These shapes are illustrated in figure 9. No nonorientable example can be put into three-dimensional Euclidean space in a clean way; any such placement is forced to have portions that pass through other portions, as can be seen in the illustration of the Klein bottle.

How does one prove that the list given above exhausts all two-dimensional manifolds? One method uses versions of the geometric techniques that are discussed below in the three-dimensional context.

2.4 Dimension 3

There is now a complete classification of all smooth, three-dimensional manifolds; however, this is a very recent achievement. There has been for some time a conjectured list of all three-dimensional manifolds, and a conjectured procedure for telling one from the other. The proof of these conjectures was recently completed by Grigori Perelman; this is a much-celebrated event in the mathematics community. The proof uses geometry about which more is said in the final part of this article. Here I shall concentrate on the classification scheme.

Before getting to the classification scheme, it is necessary to introduce the notion of a *geometric structure* on a manifold. Roughly speaking, this means a rule for defining the lengths of paths on the manifold. This rule must satisfy the following conditions. The constant path that simply stays at one point has length 0, but any path that moves at all has positive length. Second, if one path starts where another ends, the length of their concatenation (that is, the result of putting the two paths together) is the sum of their lengths.

Note that a rule of this sort for path lengths leads naturally to a notion of distance $d(x,y)$ between any two points x and y on the manifold: one takes the length of the shortest path between them. It turns out to be particularly interesting when $d(x,y)^2$ varies as a smooth function of x and y.

As it happens, there is nothing special about having a geometric structure. Manifolds have them in spades. The following are three very useful geometric structures for the interior of the ball of radius 2 about the origin in n-dimensional Euclidean space. In these formulas, the given path is viewed as if drawn in real time by some hyper-dimensional artist, with $x(t)$ denoting the position of the pencil tip on the path at time t. Here, t ranges over some interval of the real line:

$$\left.\begin{aligned} \text{length} &= \int |\dot{x}(t)|\,\mathrm{d}t; \\ \text{length} &= \int |\dot{x}(t)| \frac{1}{1+\frac{1}{4}|x(t)|^2}\,\mathrm{d}t; \\ \text{length} &= \int |\dot{x}(t)| \frac{1}{1-\frac{1}{4}|x(t)|^2}\,\mathrm{d}t. \end{aligned}\right\} \qquad (1)$$

In these formulas, $\dot{x}$ denotes the time-derivative of the path $t \to x(t)$.

The first of these geometric structures leads to the standard Euclidean distance between pairs of points. For this reason it is called the *Euclidean geometry* for the ball. The second defines what is called *spherical geometry* because the distance between any two points is the angle between certain corresponding points in the sphere of radius 1 in $(n+1)$-dimensional Euclidean space. The correspondence comes from an $(n+1)$-dimensional version of the stereographic projection that is used for maps of the Earth's polar regions. The third distance function defines what is called the *hyperbolic geometry* on the ball. This arises when the ball of radius 2 in n-dimensional Euclidean space is identified in a certain way with a particular hyperbola in $(n+1)$-dimensional Euclidean space.

The geometric structures that are depicted in (1) turn out to be symmetrical with respect to rotations and certain other transformations of the unit ball. (You can read more about Euclidean, spherical, and hyperbolic geometry in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §§6.2, 6.5, 6.6].)

As was remarked above, there are very many geometric structures on any given manifold and so one might hope to find one that has some particularly desirable properties. With this goal in mind, suppose that I have specified some "standard" geometric structure S for the ball in $\mathbb{R}^n$ to serve as a model of an exceptionally desirable structure. This could be one of the ones I have just defined or some other favorite. This leads to a corresponding notion of the structure S for a compact manifold. Roughly speaking, one says that a geometric structure on a manifold is of the type S if every point in the manifold feels as though it belongs to the unit ball with the structure S, that is, if one can use the structure S on the ball to provide coordinate charts that respect the geometric structure on the manifold. To be more precise, suppose that I am defining a coordinate system in a small neighborhood N of x by means of a function $\phi : N \to \mathbb{R}^d$. If I can always do this in such a way that the image $\phi(N)$ lies inside the ball, and such that the distance between any two points x and y in N equals the distance between their images $\phi(x)$ and $\phi(y)$, defined in terms of the structure S on the ball, then I will say that the manifold has structure of type S. In particular, a geometric structure is said to be *Euclidean*, *spherical*, or *hyperbolic* when the structure on the ball is Euclidean, spherical, or hyperbolic, respectively.

For example, the sphere in any dimension has a spherical geometric structure (as it should!). As it turns out, every two-dimensional manifold has a geometric structure that is either spherical, Euclidean, or hyperbolic. Moreover, if it has a structure of one of these types, then it cannot have one of a different type. In particular, the sphere has a spherical structure, but not a Euclidean or hyperbolic structure. Meanwhile, the torus in dimension 2 has a Euclidean geometric structure but only a Euclidean one, and all of the other manifolds listed in figure 7 have hyperbolic geometric structures and only hyperbolic ones.

William Thurston had the great insight to realize that three-dimensional manifolds might be classifiable using geometric structures. In particular, he made what was known as the *geometrization conjecture*, which says, roughly speaking, that every three-dimensional manifold is made up of "nice" pieces:

Every smooth three-dimensional manifold can be cut in a canonical fashion along a predetermined set of two-dimensional spheres and one-holed tori so that each of the resulting parts has precisely one of a list of eight possible geometric structures.

The eight possible structures include the spherical, Euclidean, and hyperbolic ones. These plus the other five are, in a sense that can be made precise, those that are maximally symmetric. The other five are associated with various LIE GROUPS [III.48 §1], as are the listed three.

Since its proof by Perelman, the geometrization conjecture has come to be known as the geometrization theorem. As I shall explain in a moment, this provides a satisfactory resolution of the three-dimensional part of the quest set out at the end of section 1. This is because a manifold with one of the eight geometric structures can be described in a canonical fashion using group theory. As a result, the geometrization theorem turns the classification issue for manifolds into a question that group theory can answer. What follows is an indication of how this comes about.

Each of the eight geometric structures has an associated *model space* which has the given geometric structure. For example, in the case of the spherical structure, the model space is the three-dimensional sphere. For the Euclidean structure, the model space is the three-dimensional Euclidean space. For the hyperbolic structure, it is the hyperbola in the four-dimensional Euclidean space, where the coordinates (x, y, z, t) obey $t^2 = 1+x^2+y^2+z^2$. In all of the eight cases, the model space has a canonical group of self-maps that preserve the distance between any two pairs of points. In the Euclidean case, this group is the group of translations and rotations of the three-dimensional Euclidean space. In the spherical case, it is the group of rotations of the four-dimensional Euclidean space, and in the hyperbolic case, it is the group of Lorentz transformations of four-dimensional Minkowski space. The associated group of self-maps is called the *isometry group* for the given geometric structure.

The connection between manifolds and group theory arises because a certain set of discrete subgroups of the isometry group of any one of the eight model spaces determines a compact manifold with the corresponding geometric structure. (A subgroup is called *discrete* if every point in the subgroup is isolated, meaning that it belongs to a neighborhood that contains no other points from the subgroup.) This compact manifold is obtained as follows. Two points x and y in the model space are declared to be *equivalent* if there is an isometry T, belonging to the subgroup, such that $Tx = y$. In other words, x is equivalent to all its images under isometries from the subgroup. It is easy to check that this notion of equivalence is a genuine EQUIVALENCE RELATION [I.2 §2.3]. The equivalence classes are then in one-to-one correspondence with the points of the associated compact manifold.

Here is a one-dimensional example of how this works. Think of the real line as a model space whose isometry group is the group of translations. The set of translations by integer multiples of 2π forms a discrete subgroup of this group. Given a point t in the real line, the possible images under translations from the subgroup are all the numbers of the form $t + 2n\pi$, where n is an integer, so one regards two real numbers as equivalent if they differ by a multiple of 2π, and the equivalence class of t is $\{t + 2n\pi : n \in \mathbb{Z}\}$. One can associate with this equivalence class the point $(x, y) = (\cos t, \sin t)$ in the circle, since adding a multiple of 2π to t does not affect either its sine or its cosine. (Intuitively speaking, if you regard each t as equivalent to $t + 2\pi$, then you are wrapping the real line around and around a circle.)

This association between certain subgroups of the isometry group and compact manifolds with the given geometric structure goes in the other direction as well. That is, the subgroup can be recovered from the manifold in a relatively straightforward fashion using the fact that each point in the manifold lies in a coordinate chart where its distance function is the same as that of the associated model space.

Even before Perelman's work there was a tremendous amount of evidence for the validity of the geometrization conjecture, much of it supplied by Thurston. In order to discuss this evidence, a small digression is required to give some of the background. First, I need to bring in the notion of a *link* in the three-dimensional sphere. A link is the name given to a finite disjoint

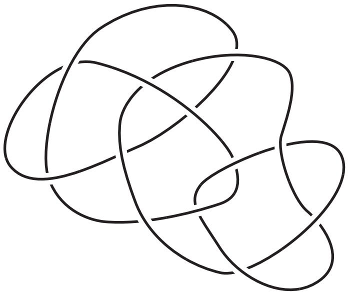

Figure 10 A link formed out of two knots.

union of knots. Figure 10 depicts an example of one that is made out of two knots.

I also need the notion of *surgery on a link*. To this end, thicken the link so as to view it as a union of knotted, solid tubes. (Think of the knot as the copper in an insulated wire and view the solid tube as the copper plus the surrounding insulation.) Notice that the boundary of any given component tube is really a copy of our one-holed torus from figure 7. Therefore, removing any one of the tubes leaves a tubular-shaped missing region from the three-dimensional sphere whose boundary is a torus.

Now, to define a surgery, imagine removing a knotted tube and then gluing it back in a different way. That is, imagine gluing the boundary of the tube to the boundary of the resulting missing region using an identification that is *not* the same as the original. For example, take the "unknot," a standard round circle in a given plane, here viewed as living inside a coordinate chart of the three-dimensional sphere. Take out the solid tube around it, and then replace the tube by gluing the boundary in the "wrong" way, as follows. Consider the leftmost torus in figure 11 as the boundary of the complement of the tube in $\mathbb{R}^3$. Consider the middle torus as the inside of the tube. The "wrong" gluing identifies the circles marked "R" and "L" on the leftmost torus with their counterparts on the middle torus. The resulting space is a three-dimensional manifold which turns out to be the product of the circle with the two-dimensional sphere. That is to say, it is the set of ordered pairs (x, y), where x is a point in the circle and y is a point in the two-dimensional sphere. There are many other possible ways to glue the boundary torus, and almost all of the corresponding surgeries give rise to distinct three-dimensional manifolds. One of these is illustrated in the rightmost part of figure 11.

In general, given any link one can construct a countably infinite set of distinct, smooth three-dimensional manifolds by using surgeries on it. Furthermore, Raymond Lickorish proved that *every* three-dimensional manifold can be obtained by using surgery on *some* link in the three-dimensional sphere. Unfortunately, this characterization of three-dimensional manifolds via surgeries on links does not provide a satisfactory resolution to the central quest of classifying smooth structures because the process is far from unique: for any given manifold there is a bewildering assortment of links and surgeries that can be used to produce it. Moreover, as of this writing, there is no known way to classify knots and links in the three-dimensional sphere.

In any event, here is a taste of Thurston's evidence for his geometrization conjecture. Given any link, all but finitely many of the three-dimensional manifolds you can produce from it by surgery satisfy the conclusions of the geometrization conjecture. Thurston also proved that, given any knot apart from the unknot, all but finitely many surgeries on it produce a manifold with a hyperbolic geometric structure.

By the way, Perelman's proof of the geometrization theorem gives as a special case a proof of the *Poincaré conjecture*, proposed by Poincaré in 1904. To state this we need the notion of a *simply connected* manifold. This is a manifold with the property that any closed loop in it can be shrunk down to a point. To be more precise, designate a point in the manifold as the "base point." Then any path in the manifold that starts and ends at the chosen base point can be continuously deformed in such a way that at each stage of the deformation the path still starts and ends at the base point, and so that the end result is the trivial path that starts at the base point and just stays there. For example, the two-dimensional sphere is simply connected, but the torus is not, since a loop that goes "once around" the torus (for example, any of the loops R or L in the various tori of figure 11) cannot be shrunk to a point. In fact, a sphere is the only two-dimensional manifold that is simply connected, and spheres are simply connected in all dimensions greater than 1.

The Poincaré conjecture. Every compact, simply connected, three-dimensional manifold is the three-dimensional sphere.

2.5 Dimension 4

This is the weird dimension. Nobody has managed to formulate a useful and viable conjecture for the classification of smooth, compact, four-dimensional manifolds. On the other hand, the classification story for

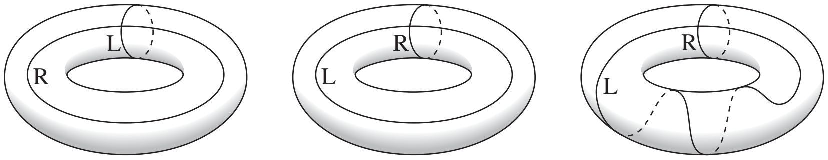

Figure 11 Different ways of gluing a tube into a tube-shaped hole.

many categories of topological four-dimensional manifolds is well-understood. For the most part, this work is by Michael Freedman.

Some of the topological manifolds in dimension 4 do not admit smooth structures. The so-called "$\frac{11}{8}$ conjecture" proposes necessary and sufficient conditions for a four-dimensional, topological manifold to have at least one smooth structure. The fraction $\frac{11}{8}$ here refers to the absolute value of the ratio of the rank to the signature of a certain symmetric, bilinear form that appears in the four-dimensional story. The case $\frac{0}{0}$ excepted, the conjecture asserts that a smooth structure exists if and only if this ratio is at least $\frac{11}{8}$. The bilinear form in question is obtained by counting with signed weights the intersection points between various two-dimensional surfaces inside the given four-dimensional manifold. In this regard, note that a typical pair of two-dimensional surfaces in four dimensions will intersect at finitely many points. This is a higher-dimensional analogue of a fact that is rather easier to visualize: that a typical pair of loops in the two-dimensional plane will intersect at finitely many points. Not surprisingly, the bilinear form here is called the *intersection form*; it plays a prominent role in Freedman's classification theorems.

Meanwhile, the problem of listing all smooth structures is wide open in four dimensions: there are no cases of a topological manifold with at least one smooth structure where the list of distinct structures is known to be complete. Some topological four-dimensional manifolds are known to have (countably) infinitely many distinct smooth structures. For others there is only one known structure. For example, the four-dimensional sphere has one obvious smooth structure and this is the only one known. However, the underlying topological manifold may, for all anyone knows, have many distinct smooth structures. By the way, the story for noncompact manifolds in dimension 4 is truly bizarre. For example, it is known that there are uncountably many smooth manifolds that are homeomorphic to the standard, four-dimensional Euclidean space. But even here, our understanding is less than optimal since there is no known explicit construction of a single one of these "exotic" smooth structures.

Simon Donaldson provided a set of geometric invariants that have the power to distinguish smooth structures on a given topological 4-manifold. Donaldson's invariants were recently superseded by a suite of more computable invariants; these were proposed by Edward Witten and are called the *Seiberg–Witten invariants*. More recently still, Peter Oszvath and Zoltan Szabo designed a possibly equivalent set of invariants that are even easier to use. Do the Seiberg–Witten invariants (broadly defined) distinguish all smooth structures? No one knows. A bit more is said about these invariants in the final part of this article.

Note that Freedman's results include the topological version of the four-dimensional Poincaré conjecture that follows.

The four-dimensional sphere is the only compact, topological 4*-manifold with the following property: every based map from either a one-dimensional circle or a two-dimensional sphere can be continuously deformed so that the result maps onto the base point.*

The smooth version of this conjecture has not been resolved.

Is there a four-dimensional version of the geometrization conjecture/theorem?

2.6 Dimensions 5 and Greater

Surprisingly enough, the issues raised at the end of the first section have more or less been resolved in all dimensions that are greater than 4. This was done some time ago by Stephen Smale with input from John Stallings. In these higher dimensions it is also possible to say what conditions need to hold in order for a topological manifold to admit a smooth structure. For example, John Milnor and others determined that the respective number of smooth structures on the spheres of dimensions 5–18 are as follows: 1, 1, 28, 2, 8, 6, 992, 1, 3, 2, 16 256, 2, 16, 16.

At first sight, it is surprising that the dimensions greater than 4 are easier to deal with than dimensions 3 and 4. However, there is a good reason for this. It turns out that there is more room to maneuver in these higher-dimensional spaces and this extra room makes all the difference. To get a sense for this, let n be a positive integer, and let S^n denote the n-dimensional sphere. To make this more concrete, view S^n as the set of points $(x_1, \dots, x_{n+1})$ in the Euclidean space $\mathbb{R}^n$ such that $x_1^2 + \cdots + x_{n+1}^2 = 1$. Now consider the product manifold, $S^n \times S^n$. This is the set of pairs of points (x, y), where x is in one copy of S^n and y is in another. This product manifold has dimension $2n$. A standard picture of $S^n \times S^n$ has two distinguished copies of S^n inside it, one consisting of all points of the form (x, y) with $y = (1, 0, \dots)$ and the other consisting of all points (x, y) with $x = (1, 0, \dots)$. Let us call the first copy S_{R} and the second one S_{L}. Of particular interest here is the fact that S_{R} and S_{L} intersect in precisely one point, the point $((1, 0, \dots), (1, 0, \dots))$.

By the way, in the $n = 1$ case, the space $S^1 \times S^1$ is the doughnut in figure 7. The one-dimensional spheres S_{R} and S_{L} inside it are the circles that are drawn in the leftmost diagram in figure 11.

If you are with me so far, suppose now that an advanced alien en route from Arcturus to the galactic center kidnaps you and drops you into some unknown, $2n$-dimensional manifold. You suspect that it is $S^n \times S^n$, but are not sure. One reason that you suspect this to be the case is that you have found a pair of n-dimensional spheres in it, one you call M_{R} and the other you call M_{L}. Unfortunately, they intersect in $2N + 1$ points, where $N > 0$. You would be less nervous about things if you could find a pair of different spheres that intersect precisely once. So you wonder whether perhaps you can push M_{L} around a bit so as to remove the $2N$ unwanted intersection points.

The surprise here is that the issue of removing intersection points in any dimension concerns only certain zero-, one-, and two-dimensional manifolds that live inside your $2n$-dimensional one. This is an old observation due to Hassler Whitney. In particular, Whitney discovered that in the $2n$-dimensional manifold you must be able to find a disk of dimension two whose boundary loop lies half in M_{L} and half in M_{R}. This boundary loop must hit two of the intersection points (one when it passes from M_{L} to M_{R} and one when it passes back again). The disk must also stick out orthogonally to M_{L} and M_{R} where it touches them. If its interior is disjoint from both M_{L} and M_{R}, and if there are no points where the disk comes back to intersect itself, then you can push the part of M_{L} that is very near the disk along the disk while stretching the remaining part to keep things from tearing. If you extend the disk a bit past M_{R}, then you will have removed two of the intersection points when you have pushed past the end of the disk. Figure 12 is a schematic of this. This pushing operation (the *Whitney trick*) can be performed in any manifold of any dimension if you can find the required disk. The problem is to find the disk. Figure 13 is a drawing of a cross-sectional slice showing a "good" disk on the left and some badly chosen disks in the middle and on the right. If you have a badly chosen disk that nevertheless satisfies the required boundary conditions, then you might hope to find a tiny wiggle of its interior that makes it better. You would like the new disk to have no self-intersection points and you would like its interior to be disjoint from both M_{L} and M_{R}. No wiggle along a direction that is parallel to the disk itself will help, for any such wiggle only changes the position of the intersection point in the disk. Likewise, a wiggle in a direction parallel to the offending M_{L} or M_{R} is useless since it only changes the position of the intersection point in the latter space. Thus, $2 + n$ of the $2n$ dimensions are useless when it comes to wiggling a disk. However, there are $2n - (n + 2) = n - 2$ remaining dimensions to work with, which is a positive number when $2n > 4$. In fact, when this is true a generic wiggle in any of these extra dimensions does the trick.

Now, when $2n = 4$ (so $n = 2$) there are no extra dimensions, and, consequently, no small wiggle can make a new disk without intersection points. So if a given candidate disk intersects M_{R}, then the Whitney trick just trades the old pair of intersection points for a new collection. If the disk intersects either itself or M_{L}, then the new version of M_{L} has self-intersection points: that is, points where one part has come around to intersect another.

This failure of the Whitney trick is the bane of four-dimensional topology. Thus, a major lemma for Michael Freedman's classification theorem about topological four-dimensional manifolds describes ubiquitous circumstances where a topologically (but not smoothly!) embedded disk can be found for use in the Whitney trick.

3 How Geometry Enters the Fray

Much of our current understanding about smooth manifolds in dimensions 4 or less has come via what

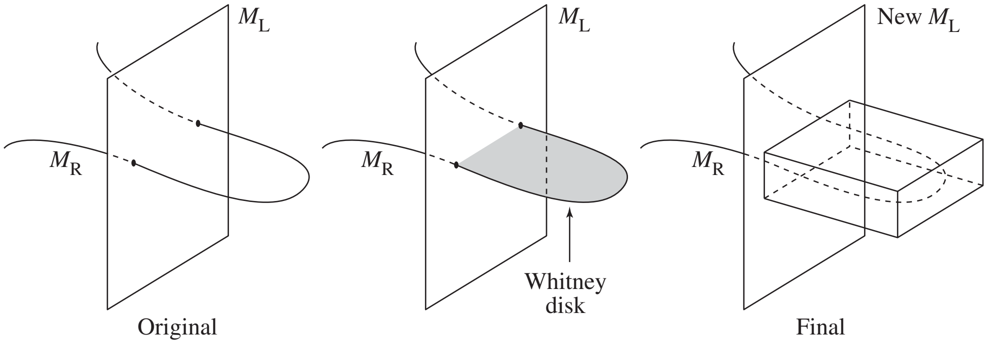

Figure 12 The Whitney trick.

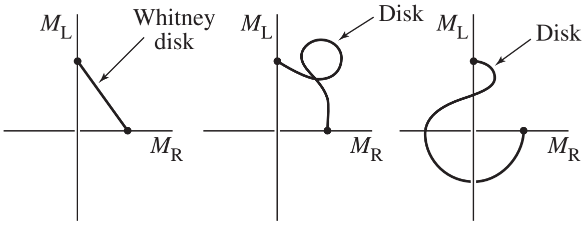

Figure 13 Some possible Whitney disks.

might be called geometric techniques. The search for a canonical geometric structure on a given three-dimensional manifold is an example. Perelman's proof of the geometrization theorem proceeds in this manner. The idea is to choose any convenient geometric structure on a given three-dimensional manifold and then continuously deform it by some well-defined rule. If one views the deformation as a time-dependent process, then the goal is to design the deformation rule to make the geometric structure ever more symmetric as time goes on.

A rule introduced and much studied by Richard Hamilton and then used by Perelman specifies the time-derivative of the geometric structure at any given time in terms of certain of its properties at that time. It is a nonlinear version of the classical HEAT EQUATION [I.3 §5.4]. For those unfamiliar with the latter, the simplest version modifies functions on the real line and will now be described. Let τ denote the time parameter, and let $f(x)$ denote a given function on the line, representing the initial distribution of heat. The resulting time-dependent family of functions associates with any given positive value for τ a function, $F_\tau(x)$, which represents the distribution of heat at time τ. The partial derivative of $F_\tau(x)$ with respect to τ is equal to its second partial derivative with respect to x, and the initial condition is that $F_0(x) = f(x)$. If the initial function f is zero outside some interval, then one can write down a formula for F_τ:

$$F_\tau(x) = \frac{1}{(2\pi\tau)^{1/2}} \int_{-\infty}^{\infty} e^{-(x-y)^2/2\tau} f(y)\,dy. \qquad (2)$$

One can see from (2) that $F_\tau(x)$ tends uniformly to zero in x as τ tends to infinity. In particular, this limit is completely ignorant of the starting function f; and, being identically zero, it is also the most symmetric function possible. The representation for F_τ in (2) indicates how this comes about. The value of F_τ at any given point is a weighted average of the values of the original function. Moreover, as τ increases, this average looks more like the standard average over ever-larger regions of the line. Physically this is very plausible as well: the heat spreads itself out more and more thinly as time goes on.

The time-dependent family of geometric structures that Hamilton introduced and Perelman used is defined by an equation that relates the time-derivative of the geometric structure at any given time to its *Ricci curvature*, a certain natural substitute in the context of geometric structures for the second derivatives that enter the heat equation for the functions F_τ above. The idea much studied by Hamilton and then by Perelman is to let the evolving geometric structure decompose the manifold into the canonical pieces that are predicted to exist by the geometrization conjecture. Perelman proved that the pieces required by the geometrization conjecture emerge as regions whose points stay relatively close together (as measured by a certain rescaling of the distance function) while the points in distinct regions move farther and farther apart.

The equation used by Perelman and Hamilton for the time-evolution of a geometric structure is rather complicated. Its standard incarnation involves the notion of a RIEMANNIAN METRIC [I.3 §6.10]. This appears in any given coordinate chart on an n-dimensional manifold as a symmetric, positive-definite $n \times n$ matrix whose entries are functions of the coordinates. The various components of this matrix are traditionally written as $\{g_{ij}\}_{1 \leqslant i, j \leqslant n}$. The matrix determines the geometric structure and can in turn be derived from it.

Hamilton and Perelman study a time-dependent family of Riemannian metrics, $\tau \to g_\tau$, where the rule for the time dependence is obtained using an equation for the τ-derivative of g_τ that has the schematic form $\partial_\tau (g_\tau)_{ij} = -2R_{ij}[g_\tau]$, where $\{R_{ij}\}_{1 \leqslant i,\, j \leqslant n}$ are the components of the aforementioned Ricci curvature, a certain symmetric matrix that is determined at any given τ by the metric g_τ. Every Riemannian metric has a Ricci curvature; its components are standard (nonlinear) functions of the components of the matrix and their first- and second-order partial derivatives in the coordinate directions. The Ricci curvatures for the metrics that define the respective Euclidean, spherical, and hyperbolic geometries have the particularly simple form $R_{ij} = cg_{ij}$, where c is 0, 1, or -1, respectively. For more about these ideas, see RICCI FLOW [III.78].

As was mentioned at the beginning of this part of the article, geometry has also played a central role in the developments in the classification program for smooth, four-dimensional manifolds. In this case, geometrically defined data are used to distinguish smooth structures on topologically equivalent manifolds. What follows is a very brief sketch of how this is done.

To begin with, the idea is to introduce a geometric structure on the manifold and then to use the latter to define a canonical system of partial differential equations. In any given coordinate chart, these equations are for a particular set of functions. The equations state that certain linear combinations of the collection of first derivatives of the functions from the set are equal to terms that are linear and quadratic in the values of the functions themselves. In the case of the Donaldson invariants, and also of the newer Seiberg–Witten invariants, the relevant equations are nonlinear generalizations of the MAXWELL EQUATIONS [IV.13 §1.1] for electricity and magnetism.

In any event, one then counts the solutions with algebraic weights. The purpose of the algebraic weighting of the count is to obtain an INVARIANT [I.4 §2.2], that is, a count that does not change if the given geometric structure is changed. The point here is that the naive count will typically depend on the structure, but a suitably weighted count will not. Imagine, for example, that one has a continuously varying family of geometric structures, and that new solutions appear and old ones disappear only in pairs, where one solution has been assigned weight $+1$ and the other -1.

The following toy model illustrates this appearance and disappearance phenomenon. The equation in question is for a single function on the circle. That is, it will concern a function, f, of one variable, x, that is periodic with period 2π. For example, take the equation $\partial f / \partial x + \tau f - f^3 = 0$, where τ is a constant that is specified in advance. Varying τ can now be viewed as a model for the variation of the geometric structure. When $\tau > 0$ there are exactly three solutions: $f \equiv 0$, $f \equiv \tau$, and $f \equiv -\tau$. However, when $\tau \leqslant 0$, the only solution is $f \equiv 0$. Thus, the number of solutions changes as τ crosses zero. Even so, a suitable weighted count is independent of τ.

Let us return now to the four-dimensional story. If the weighted sum is independent of the chosen *geometric* structure, then it depends only on the underlying *smooth* structure. Therefore, if two geometric structures on a given topological manifold provide distinct sums, then the underlying smooth structures must be distinct.

As I remarked earlier, Oszvath and Szabo have defined invariants for four-dimensional manifolds that are easier to use than the Seiberg–Witten invariants, but probably equivalent to them. These are also defined as the number of solutions to a particular system of differential equations, counted in a creative way. In this case, the equations are analogues of the CAUCHY–RIEMANN EQUATIONS [I.3 §5.6], and the arena is a space that can be defined after cutting the 4-manifold into simpler pieces. There are myriad ways to slice a 4-manifold in the prescribed manner, but a suitably creative, algebraic count of solutions provides the same number for each.

With hindsight, one can see that the use of differential equations to distinguish smooth structures on a given topological manifold makes good sense, since a smooth structure is needed to take a derivative in the first place. Even so, this author is constantly amazed by the fact that the Donaldson/Seiberg–Witten/Oszvath–Szabo strategy of algebraically counting differential equation solutions yields counts that are both tractable and useful. (Getting the same count in all cases is no help at all.)

Further Reading

Those who wish to learn more about manifolds in general can consult J. Milnor's book *Topology from the Differentiable Viewpoint* (Princeton University Press, Princeton, NJ, 1997) or the book *Differential Topology* (Prentice Hall, Englewood Cliffs, NJ, 1974), by V. Guillemin and A. Pollack. A good introduction to the classification problem in dimensions 2 and 3 is the book *Three-Dimensional Geometry and Topology* (Princeton University Press, Princeton, NJ, 1997), by W. Thurston. This book also has a nice discussion of geometric structures. A full account of Perelman's proof of the Poincaré conjecture can be found in *Ricci Flow and the Poincaré Conjecture*, by J. Morgan and G. Tian (American Mathematical Society, Providence, RI, 2007). The story for topological 4-manifolds is told in the book by M. Freedman and F. Quinn titled *Topology of* 4*-Manifolds* (Princeton University Press, Princeton, NJ, 1990). There are no books available that serve as general introductions to the smooth 4-manifold story. A book that does introduce the Seiberg–Witten invariants is *The Seiberg–Witten Equations and Applications to the Topology of Smooth Four-Manifolds* (Princeton University Press, Princeton, NJ, 1995), by J. Morgan. Meanwhile, the Donaldson invariants are discussed in detail in the book by Donaldson and P. Kronheimer titled *Geometry of Four-Manifolds* (Oxford University Press, Oxford, 1990). Finally, parts of the story for dimensions greater than 4 are told in *Lectures on the h-Cobordism Theorem* (Princeton University Press, Princeton, NJ, 1965), by J. Milnor, and *Foundational Essays on Topological Manifolds, Smoothings and Triangulations* (Princeton University Press, Princeton, NJ, 1977), by R. Kirby and L. Siebenman.

IV.8 Moduli Spaces

David D. Ben-Zvi

Many of the most important problems in mathematics concern CLASSIFICATION [I.4 §2]. One has a class of mathematical objects and a notion of when two objects should count as equivalent. It may well be that two equivalent objects look superficially very different, so one wishes to describe them in such a way that equivalent objects have the same description and inequivalent objects have different descriptions.

Moduli spaces can be thought of as *geometric* solutions to *geometric* classification problems. In this article we shall illustrate some of the key features of moduli spaces, with an emphasis on the moduli spaces of RIEMANN SURFACES [III.79]. In broad terms, a *moduli problem* consists of three ingredients.

Objects: which geometric objects would we like to describe, or *parametrize*?

Equivalences: when do we identify two of our objects as being isomorphic, or "the same"?

Families: how do we allow our objects to vary, or modulate?

In this article we will discuss what these ingredients signify, as well as what it means to *solve* a moduli problem, and we will give some indications as to why this might be a good thing to do.

Moduli spaces arise throughout ALGEBRAIC GEOMETRY [IV.4], differential geometry, and ALGEBRAIC TOPOLOGY [IV.6]. (Moduli spaces in topology are often referred to as *classifying spaces*.) The basic idea is to give a geometric structure to the *totality* of the objects we are trying to classify. If we can understand this geometric structure, then we obtain powerful insights into the geometry of the objects themselves. Furthermore, moduli spaces are rich geometric objects in their own right. They are "meaningful" spaces, in that any statement about their geometry has a "modular" interpretation, in terms of the original classification problem. As a result, when one investigates them one can often reach much further than one can with other spaces. Moduli spaces such as the moduli of ELLIPTIC CURVES [III.21] (which we discuss below) play a central role in a variety of areas that have no immediate link to the geometry being classified, in particular in ALGEBRAIC NUMBER THEORY [IV.1] and algebraic topology. Moreover, the study of moduli spaces has benefited tremendously in recent years from interactions with physics (in particular with STRING THEORY [IV.17 §2]). These interactions have led to a variety of new questions and new techniques.

1 Warmup: The Moduli Space of Lines in the Plane

Let us begin with a problem that looks rather simple, but that nevertheless illustrates many of the important ideas of moduli spaces.

Problem. Describe the collection of all lines in the real plane $\mathbb{R}^2$ that pass through the origin.

To save writing, we are using the word "line" to mean "line that passes through the origin." This classification compactifying moduli spaces is that we can then calculate integrals over the completed space. This is crucial for the next item.

Invariants from moduli spaces. An important application of moduli spaces in geometry and topology is inspired by quantum field theory, where a particle, rather than following the "best" classical path between two points, follows all paths with varying probabilities (see MIRROR SYMMETRY [IV.16 §2.2.4]). Classically, one calculates many topological invariants by picking a geometric structure (such as a metric) on a space, calculating some quantity using this structure, and finally proving that the result of the calculation did not depend on the structure we chose. The new alternative is to look at *all* such geometric structures, and integrate some quantity over the space of all choices. The result, if we can show convergence, will manifestly not depend on any choices. String theory has given rise to many important applications of this idea, in particular by giving a rich structure to the collection of integrals obtained in this way. Donaldson and Seiberg–Witten theories use this philosophy to give topological invariants of four-manifolds. Gromov–Witten theory applies it to the topology of SYMPLECTIC MANIFOLDS [III.88], and to counting problems in algebraic geometry, such as, How many rational plane curves of degree 5 pass through fourteen points in general position? (Answer: 87 304.)

Modular forms. One of the most profound ideas in mathematics, the Langlands program, relates number theory to function theory (harmonic analysis) on very special moduli spaces, generalizing the moduli space of elliptic curves. These moduli spaces (Shimura varieties) are expressible as quotients of symmetric spaces (such as $\mathbb{H}$) by arithmetic groups (such as $PSL_2(\mathbb{Z})$). MODULAR FORMS [III.59] and automorphic forms are special functions on these moduli spaces, described by their interaction with the large symmetry groups of the spaces. This is an extremely exciting and active area of mathematics, which counts among its recent triumphs the proof of FERMAT'S LAST THEOREM [V.10] and the Shimura–Taniyama–Weil conjecture (Wiles, Taylor–Wiles, Breuil–Conrad–Diamond–Taylor).

Further Reading

For historical accounts and bibliographies on moduli spaces, the following articles are highly recommended.

A beautiful and accessible overview of moduli spaces, with an emphasis on the notion of deformations, is given by Mazur (2004). The articles by Hain (2000) and Looijenga (2000) give excellent introductions to the study of the moduli spaces of curves, perhaps the oldest and most important of all moduli problems. The article by Mumford and Suominen (1972) introduces the key ideas underlying the study of moduli spaces in algebraic geometry.

Hain, R. 2000. Moduli of Riemann surfaces, transcendental aspects. In *School on Algebraic Geometry, Trieste, 1999*, pp. 293–353. ICTP Lecture Notes Series, no. 1. Trieste: The Abdus Salam International Centre for Theoretical Physics.

Looijenga, E. 2000. A minicourse on moduli of curves. In *School on Algebraic Geometry, Trieste, 1999*, pp. 267–91. ICTP Lecture Notes Series, no. 1. Trieste: The Abdus Salam International Centre for Theoretical Physics.

Mazur, B. 2004. Perturbations, deformations and variations (and "near-misses") in geometry. Physics and number theory. *Bulletin of the American Mathematical Society* 41(3):307–36.

Mumford, D., and K. Suominen. 1972. Introduction to the theory of moduli. In *Algebraic Geometry, Oslo, 1970: Proceedings of the Fifth Nordic Summer School in Mathematics*, edited by F. Oort, pp. 171–222. Groningen: Wolters-Noordhoff.

IV.9 Representation Theory

Ian Grojnowski

1 Introduction

It is a fundamental theme in mathematics that many objects, both mathematical and physical, have symmetries. The goal of GROUP [I.3 §2.1] theory in general, and representation theory in particular, is to study these symmetries. The difference between representation theory and general group theory is that in representation theory one restricts one's attention to symmetries of VECTOR SPACES [I.3 §2.3]. I will attempt here to explain why this is sensible and how it influences our study of groups, causing us to focus on groups with certain nice structures involving *conjugacy classes*.

2 Why Vector Spaces?

The aim of representation theory is to understand how the *internal* structure of a group controls the way it acts *externally* as a collection of symmetries. In the other direction, it also studies what one can learn about a group's internal structure by regarding it as a group of symmetries.

We begin our discussion by making more precise what we mean by "acts as a collection of symmetries." The idea we are trying to capture is that if we are given a group G and an object X, then we can associate with each element g of G some symmetry of X, which we call $\phi(g)$. For this to be sensible, we need the composition of symmetries to work properly: that is, $\phi(g)\phi(h)$ (the result of applying $\phi(h)$ and then $\phi(g)$) should be the same symmetry as $\phi(gh)$. If X is a set, then a symmetry of X is a particular kind of PERMUTATION [III.68] of its elements. Let us denote by $\operatorname{Aut}(X)$ the group of *all* permutations of X. Then an *action* of G on X is defined to be a homomorphism from G to $\operatorname{Aut}(X)$. If we are given such a homomorphism, then we say that G *acts* on X.

The image to have in mind is that G "does things" to X. This idea can often be expressed more conveniently and vividly by forgetting about ϕ in the notation: thus, instead of writing $\phi(g)(x)$ for the effect on x of the symmetry associated with g, we simply think of g itself as a permutation and write gx. However, sometimes we do need to talk about ϕ as well: for instance, we might wish to compare two different actions of G on X.

Here is an example. Take as our object X a square in the plane, centered at the origin, and let its vertices be A, B, C, and D (see figure 1). A square has eight symmetries: four rotations by multiples of 90° and four reflections. Let G be the group consisting of these eight symmetries; this group is often called D_8, or the *dihedral group* of order 8. By definition, G acts on the square. But it also acts on the set of *vertices* of the square: for instance, the action of the reflection through the y-axis is to switch A with B and C with D. It might seem as though we have done very little here. After all, we defined G as a group of symmetries so it does not take much effort to associate a symmetry with each element of G. However, we did not define G as a group of permutations of the set $\{A, B, C, D\}$, so we have at least done something.

To make this point clearer, let us look at some other sets on which G acts, which will include any set that we can build sufficiently naturally from the square. For instance, G acts not only on the set of vertices $\{A, B, C, D\}$, but on the set of edges $\{AB, BC, CD, DA\}$ and on the set of cross-diagonals $\{AC, BD\}$ as well. Notice in the latter case that some of the elements of G act in the same way: for example, a clockwise rotation through 90° interchanges the two diagonals, as does a counterclockwise rotation through 90°. If all the elements of G act differently, then the action is called *faithful*.

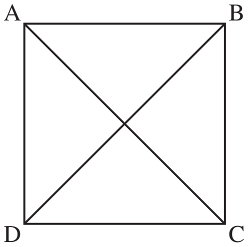

Figure 1 A square and its diagonals.

Notice that the operations on the square ("reflect through the y-axis," "rotate through 90°," and so on) can be applied to the whole Cartesian plane $\mathbb{R}^2$. Therefore, $\mathbb{R}^2$ is another (and much larger) set on which G acts. To call $\mathbb{R}^2$ a set, though, is to forget the very interesting fact that the elements in $\mathbb{R}^2$ can be added together and multiplied by real numbers: in other words, $\mathbb{R}^2$ is a *vector space*. Furthermore, the action of G is well-behaved with respect to this extra structure. For instance, if g is one of our symmetries and v_1 and v_2 are two elements of $\mathbb{R}^2$, then g applied to the sum $v_1 + v_2$ yields the sum $g(v_1) + g(v_2)$. Because of this, we say that G acts *linearly* on the vector space $\mathbb{R}^2$. When V is a vector space, we denote by $\operatorname{GL}(V)$ the set of invertible linear maps from V to V. If V is the vector space $\mathbb{R}^n$, this group is the familiar group $\operatorname{GL}_n(\mathbb{R})$ of invertible $n \times n$ matrices with real entries; similarly, when $V = \mathbb{C}^n$ it is the group of invertible matrices with complex entries.

Definition. A *representation* of a group G on a vector space V is a homomorphism from G to $\operatorname{GL}(V)$.

In other words, a group action is a way of regarding a group as a collection of permutations, while a representation is the special case where these permutations are invertible linear maps. One sometimes sees representations referred to, for emphasis, as *linear* representations. In the representation of D_8 on $\mathbb{R}^2$ that we described above, the homomorphism from G to $\operatorname{GL}_2(\mathbb{R})$ took the symmetry "clockwise rotation through 90°" to the matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and the symmetry "reflection through the y-axis" to the matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

Given one representation of G, we can produce others using natural constructions from linear algebra. For example, if ρ is the representation of G on $\mathbb{R}^2$ described above, then its DETERMINANT [III.15] $\det \rho$ is a homomorphism from G to $\mathbb{R}^*$ (the group of nonzero real numbers under multiplication), since

$$\det(\rho(gh)) = \det(\rho(g)\rho(h)) = \det(\rho(g))\det(\rho(h)),$$

by the multiplicative property of determinants. This makes $\det \rho$ a one-dimensional representation, since each nonzero real number t can be thought of as the element "multiply by t" of $\mathrm{GL}_1(\mathbb{R})$. If ρ is the representation of D_8 just discussed, then under $\det \rho$ we find that rotations act as the identity and reflections act as multiplication by -1.

The definition of "representation" is formally very similar to the definition of "action," and indeed, since every linear automorphism of V is a permutation on the set of vectors in V, the representations of G on V form a subset of the actions of G on V. But the set of representations is in general a much more interesting object. We see here an instance of a general principle: if a set comes equipped with some extra structure (as a vector space comes with the ability to add elements together), then it is a mistake not to make use of that structure; and the more structure the better.

In order to emphasize this point, and to place representations in a very favorable light, let us start by considering the general story of actions of groups on sets. Suppose, then, that G is a group that acts on a set X. For each x, the set of all elements of the form gx, as g ranges over G, is called the *orbit* of x. It is not hard to show that the orbits form a partition of X.

Example. Let G be the dihedral group D_8 acting on the set X of *ordered pairs* of vertices of the square, of which there are sixteen. Then there are three orbits of G on X, namely $\{\mathrm{AA}, \mathrm{BB}, \mathrm{CC}, \mathrm{DD}\}$, $\{\mathrm{AB}, \mathrm{BA}, \mathrm{BC}, \mathrm{CB}, \mathrm{CD}, \mathrm{DC}, \mathrm{DA}, \mathrm{AD}\}$, and $\{\mathrm{AC}, \mathrm{CA}, \mathrm{BD}, \mathrm{DB}\}$.

An action of G on X is called *transitive* if there is just one orbit. In other words, it is transitive if for every x and y in X you can find an element g such that $gx = y$. When an action is *not* transitive, we can consider the action of G on each orbit separately, which effectively breaks up the action into a collection of transitive actions on disjoint sets. So in order to study *all* actions of G on sets it suffices to study *transitive* actions; you can think of actions as "molecules" and transitive actions as the "atoms" into which they can be decomposed. We shall see that this idea of *decomposing into objects that cannot be further decomposed* is fundamental to representation theory.

What are the possible transitive actions? A rich source of such actions comes from subgroups H of G. Given a subgroup H of G, a *left coset* of H is a set of the form $\{gh : h \in H\}$, which is commonly denoted by gH. An elementary result in group theory is that the left cosets form a partition of G (as do the right cosets, if you prefer them). There is an obvious action of G on the set of left cosets of H, which we denote by G/H: if g' is an element of G, then it sends the coset gH to the coset $(g'g)H$.

It turns out that every transitive action is of this form! Given a transitive action of G on a set X, choose some $x \in X$ and let H_x be the subgroup of G consisting of all elements h such that $hx = x$. (This set is called the *stabilizer* of x.) Then one can check that the action of G on X is the same[1] as that of G on the left cosets of H_x. For example, the action of D_8 on the first orbit above is isomorphic to the action on the left cosets of the two-element subgroup H generated by a reflection of the square through its diagonal. If we had made a different choice of x, for example the point $x' = gx$, then the subgroup of G fixing x' would just be gH_xg^{-1}. This is a so-called *conjugate subgroup*, and it gives a different description of the same orbit, this time as left cosets of gH_xg^{-1}.

It follows that there is a one-to-one correspondence between transitive actions of G and conjugacy classes of subgroups (that is, collections of subgroups conjugate to some given subgroup). If G acts on our original set X in a nontransitive way, then we can break X up into a union of orbits, each of which, as a result of this correspondence, is associated with a conjugacy class of subgroups. This gives us a convenient "bookkeeping" mechanism for describing the action of G on X: just keep track of how many times each conjugacy class of subgroups arises.

Exercise. Check that in the example earlier the three orbits correspond (respectively) to a two-element subgroup R generated by reflection through a diagonal, the trivial subgroup, and another copy of the group R.

This completely solves the problem of how groups act on sets. The internal structure that controls the action is the *subgroup* structure of G.

In a moment we will see the corresponding solution to the problem of how groups act on vector spaces. First, let us just stare at sets for a while and see why, though we have answered our question, we should not feel too happy about it.[2]

The problem is that the subgroup structure of a group is *just horrible.*

[1] By "the same" we mean "isomorphic as sets with G-action." The casual reader may read this as "the same," while the more careful reader should stop here and work out, or look up, precisely what is meant.

[2] Exercise: go back to the example of D_8 and list all the possible transitive actions.

For example, any finite group of order n is a subgroup of the SYMMETRIC GROUP [III.68] S_n (this is "Cayley's theorem," which follows by considering the action of G on itself), so in order to list the conjugacy classes of subgroups of the symmetric group S_n one must understand all finite groups of size less than n.[3] Or consider the cyclic group $\mathbb{Z}/n\mathbb{Z}$. The subgroups correspond to the divisors of n, a subtle property of n that makes the cyclic groups behave quite differently as n varies. If n is prime, then there are very few subgroups, while if n is a power of 2 there are quite a few. So number theory is involved even if all we want to do is understand the subgroup structure of a group as simple as a cyclic group.

With some relief we now turn our attention back to linear representations. We will see that, just as with actions on sets, one can decompose representations into "atomic" ones. But, by contrast with the case of sets, these atomic representations (called "irreducible" representations, or sometimes simply "irreducibles") turn out to exhibit quite beautiful regularities.

The nice properties of representation theory come largely from the following fact. While elements of the symmetric group S_n can be multiplied together, elements of $\mathrm{GL}(V)$, being matrices, can be *added* as well as multiplied. (But beware: the sum of two elements of $\mathrm{GL}(V)$ is not necessarily an element of $\mathrm{GL}(V)$, because it may not be invertible. It is, however, an element of the endomorphism algebra $\mathrm{End}(V)$. When $V = \mathbb{C}^n$, $\mathrm{End}(V)$ is just the familiar algebra of all $n \times n$ matrices with complex entries, both invertible and not.)

To see the difference it makes to be able to add, consider the cyclic group $G = \mathbb{Z}/n\mathbb{Z}$. For each $\omega \in \mathbb{C}$ with $\omega^n = 1$, we get a representation χ_ω of G on $\mathbb{C}$ by associating the element $r \in \mathbb{Z}/n\mathbb{Z}$ with multiplication by ω^r, which we think of as a linear map from the one-dimensional space $\mathbb{C}$ to itself. This gives us n different one-dimensional representations, one for each nth root of unity, and it turns out that there are no others. Moreover, if $\rho : G \to \mathrm{GL}(V)$ is any representation of $\mathbb{Z}/n\mathbb{Z}$, then we can write it as a direct sum of these representations by imitating the formula for finding the Fourier mode of a function. Using the representation ρ, we associate with each r in $\mathbb{Z}/n\mathbb{Z}$ a linear map $\rho(r)$. Now let us define a linear map $p_\omega : V \to V$ by the formula

$$p_\omega = \frac{1}{n} \sum_{0 \leqslant r < n} \omega^{-r} \rho(r).$$

Then p_ω is an element of $\mathrm{End}(V)$, and one can check that it is actually a PROJECTION [III.50 §3.5] onto a subspace V_ω of V. In fact, this subspace is an EIGENSPACE [I.3 §4.3]: it consists of all vectors v such that $\rho(1)v = \omega v$, which implies, since ρ is a representation, that $\rho(r)v = \omega^r v$. The projection p_ω should be thought of as the analogue of the nth FOURIER COEFFICIENT [III.27] $a_n(f)$ of a function $f(\theta)$ on the circle; note the formal similarity of the above formula to the Fourier expansion formula $a_n(f) = \int \mathrm{e}^{-2\pi i n\theta} f(\theta)\,\mathrm{d}\theta$.

Now the interesting thing about the Fourier series of f is that, under favorable circumstances, it adds up to f itself: that is, it decomposes f into TRIGONOMETRIC FUNCTIONS [III.92]. Similarly, what is interesting about the subspaces V_ω is that we can use them to decompose the representation ρ. The composition of any two distinct projections p_ω is 0, from which it can be shown that

$$V = \bigoplus_\omega V_\omega.$$

We can write each subspace V_ω as a sum of one-dimensional spaces, which are copies of $\mathbb{C}$, and the restriction of ρ to any one of these is just the simple representation χ_ω defined earlier. Thus, ρ has been decomposed as a combination of very simple "atoms" χ_ω.[4]

This ability to add matrices has a very useful consequence. Let a finite group G act on a complex vector space V. A subspace W of V is called *G-invariant* if $gW = W$ for every $g \in G$. Let W be a G-invariant subspace, and let U be a complementary subspace (that is, one such that every element v of V can be written in exactly one way as $w + u$ with $w \in W$ and $u \in U$). Let ϕ be an arbitrary projection onto U. Then it is a simple exercise to show that the linear map $1/|G| \sum_{g \in G} g\phi$ is also a projection onto a complementary subspace, but with the added advantage that it is G-invariant. This latter fact follows because applying an element g' to the sum just rearranges its terms.

The reason this is so useful is that it allows us to decompose an arbitrary representation into a direct sum of *irreducible representations*, which are representations without a G-invariant subspace. Indeed, if ρ is

[3] THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.7] does at least allow us to estimate the *number* γ_n of subgroups of S_n up to conjugacy: it is a result of Pyber that $2^{((1/16)+o(1))n^2} \leqslant \gamma_n \leqslant 24^{((1/6)+o(1))n^2}$. Equality is expected for the lower bound.

[4] To summarize the rest of this article: the similarity to the Fourier transform is not just analogy—decomposing a representation into its irreducible summands is a notion that includes both this example and the Fourier transform.

not irreducible, then there is a G-invariant subspace W. By the above remark, we can write $G = W \oplus W'$ with W' also G-invariant. If either W or W' has a further G-invariant subspace, then we can decompose it further, and so on. We have just seen this done for the cyclic group: in that case the irreducible representations were the one-dimensional representations χ_ω.

The irreducible representations are the basic building blocks of arbitrary complex representations, just as the basic building blocks for actions on sets are the transitive actions. It raises the question of what the irreducible representations are, a question that has been answered for many important examples, but which is not yet solvable by any general procedure.

To return to the difference between actions and representations, another important observation is that any action of a group G on a finite set X can be *linearized* in the following sense. If X has n elements, then we can look at the HILBERT SPACE [III.37] $L^2(X)$ of all complex-valued functions defined on X. This has a natural basis given by the "delta functions" δ_x, which send x to 1 and all other elements of X to 0. Now we can turn the action of G on X into an action of G on the basis in an obvious way: we just define $g\delta_x$ to be δ_{gx}. We can extend this definition by linearity, since an arbitrary function f is a linear combination of the basis functions δ_x. This gives us an action of G on $L^2(X)$, which can be defined by a simple formula: if f is a function in $L^2(X)$, then gf is the function defined by $(gf)(x) = f(g^{-1}x)$. Equivalently, gf does to gx what f does to x. Thus, an action on sets can be thought of as an assignment of a very special matrix to every group element, namely a matrix with only 0s and 1s and precisely one 1 in each row and each column. (Such matrices are called *permutation matrices.*) By contrast, a general representation assigns an *arbitrary* invertible matrix.

Now, even when X itself is a single orbit under the action of G, the above representation on $L^2(X)$ can break up into pieces. For an extreme example of this phenomenon, consider the action of $\mathbb{Z}/n\mathbb{Z}$ on itself by multiplication. We have just seen that, by means of the "Fourier expansion" above, this breaks up into a sum of n one-dimensional representations.

Let us now consider the action of an arbitrary group G on itself by multiplication, or, to be more precise, left multiplication. That is, we shall associate with each element g the permutation of G that takes each h in G to gh. This action is obviously transitive. As an action on a *set* it cannot be decomposed any further. But when we *linearize* this action to a representation of G on the vector space $L^2(G)$, we have much greater flexibility to decompose the action. It turns out that, not only does it break up into a direct sum of many irreducible representations, but *every* irreducible representation ρ of G occurs as one of the summands in this direct sum, and the number of times that ρ appears is equal to the dimension of the subspace on which it acts.

The representation we have just discussed is called the *left regular representation* of G. The fact that every irreducible representation occurs in it so regularly makes it extremely useful. Notice that it is easier to decompose representations on complex vector spaces than on real vector spaces, since every automorphism of a complex vector space has an eigenvector. So it is simplest to begin by studying complex representations.

The time has now come to state the fundamental theorem about complex representations of finite groups. This theorem tells us how many irreducible representations there are for a finite group, and, more colorfully, that representation theory is a "non-Abelian analogue of Fourier decomposition."

Let $\rho : G \to \mathrm{End}(V)$ be a representation of G. The *character* χ_ρ of ρ is defined to be its trace: that is, χ_ρ is a function from G to $\mathbb{C}$ and $\chi_\rho(g) = \mathrm{tr}(\rho(g))$ for each g in G. Since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any two matrices A and B, we have $\chi_\rho(hgh^{-1}) = \chi_\rho(g)$. Therefore, χ_V is very far from an arbitrary function on G: it is a function that is constant on each *conjugacy class*. Let K_G denote the vector space of all complex-valued functions on G with this property; it is called the *representation ring* of G.

The characters of the irreducible representations of a group form a very important set of data about the group, which it is natural to organize into a matrix. The columns are indexed by the conjugacy classes, the rows by the irreducible representations, and each entry is the value of the character of the given representation at the given conjugacy class. This array is called the *character table* of the group, and it contains all the important information about representations of the group: it is our periodic table. The basic theorem of the subject is that this array is a *square*.

Theorem (the character table is square). *Let G be a finite group. Then the characters of the irreducible representations form an orthonormal basis of K_G.*

When we say that the basis of characters is *orthonormal* we mean that the Hermitian inner product defined by

$$\langle \chi, \psi \rangle = |G|^{-1} \sum_{g \in G} \chi(g)\overline{\psi(g)}$$

is 1 when $\chi = \psi$ and 0 otherwise. The fact that it is a basis implies in particular that there are exactly as many irreducible representations as there are conjugacy classes in G, and the map from isomorphism classes of representations to K_G that sends each ρ to its character is an injection. That is, an arbitrary representation is determined up to isomorphism by its character.

The internal structure of a group G that controls how it can act on vector spaces is the structure of conjugacy classes of elements of G. This is a much gentler structure than the set of all conjugacy classes of *subgroups* of G. For example, in the symmetric group S_n two permutations belong to the same conjugacy class if and only if they have the same cycle type. Therefore, in that group there is a bijection between conjugacy classes and partitions of n.[5]

Furthermore, whereas it is completely unclear how to count subgroups, conjugacy classes are much easier to handle. For instance, since they partition the group, we have the formula $|G| = \sum_{C \text{ a conjugacy class}} |C|$. On the representation side, there is a similar formula, which arises from the decomposition of the regular representation $L^2(G)$ into irreducibles: $|G| = \sum_{V \text{ irreducible}} (\dim V)^2$. It is inconceivable that there might be a similarly simple formula for sums over all subgroups of a group.

We have reduced the problem of understanding the general structure of the representations of a finite group G to the problem of determining the character table of G. When $G = \mathbb{Z}/n\mathbb{Z}$, our description of the n irreducible representations above implies that all the entries of this matrix are roots of unity. Here are the character tables for D_8 (on the left), the group of symmetries of the square, and, just for contrast, for the group $\mathbb{Z}/3\mathbb{Z}$ (on the right):

1	1	1	1	1
1	1	1	−1	−1
1	1	−1	1	−1
1	1	−1	−1	1
2	−2	0	0	0

1	1	1
1	z	z^2
1	z^2	z

where $z = \exp(2\pi\mathrm{i}/3)$.

The obvious question—Where did the first table come from?—indicates the main problem with the theorem: though it tells us the shape of the character table, it leaves us no closer to understanding what the actual character values are. We know *how many* representations there are, but not *what* they are, or even what their dimensions are. We do not have a general method for constructing them, a kind of "non-Abelian Fourier transform." This is the central problem of representation theory.

Let us see how this problem can be solved for the group D_8. Over the course of this article, we have already encountered three irreducible representations of this group. The first is the "trivial" one-dimensional representation: the homomorphism $\rho : D_8 \to \mathrm{GL}_1$ that takes every element of D_8 to the identity. The second is the two-dimensional representation we wrote down in the first section, where each element of D_8 acts on $\mathbb{R}^2$ in the obvious way. The determinant of this representation is a one-dimensional representation that is *not* trivial: it sends the rotations to 1 and the reflections to -1. So we have constructed three rows of the character table above. There are five conjugacy classes in D_8 (trivial, reflection through axis, reflection through diagonal, 90° rotation, 180° rotation), so we know that there are just two more rows.

The equality $|G| = 8 = 2^2 + 1 + 1 + (\dim V_4)^2 + (\dim V_5)^2$ implies that these missing representations are one dimensional. One way of getting the missing character values is to use orthogonality of characters.

A slightly (but only slightly) less ad hoc way is to decompose $L^2(X)$ for small X. For example when X is the pair of diagonals $\{\mathrm{AC}, \mathrm{BD}\}$, we have $L^2(X) = V_4 \oplus \mathbb{C}$, where $\mathbb{C}$ is the trivial representation.

We are now going to start pointing the way toward some more modern topics in representation theory. Of necessity, we will use language from fairly advanced mathematics: the reader who is familiar with only some of this language should consider browsing the remaining sections, since different discussions have different prerequisites.

In general, a good, but not systematic, way of finding representations is to find objects on which G acts, and "linearize" the action. We have seen one example of this: when G acts on a set X we can consider the linearized action on $L^2(X)$. Recall that the irreducible G-sets are all of the form G/H, for H some subgroup of G. As well as looking at $L^2(G/H)$, we can consider, for every representation W of H, the vector space $L^2(G/H, W) = \{f : G \to W \mid f(gh) = h^{-1}f(g),\ g \in G,\ h \in H\}$; in geometric language, for those who prefer it, this is the space of sections of the associated W-bundle on G/H. This representation of G is called the *induced representation* of W from H to G.

[5] Not only is the set of all partitions a sensible combinatorial object, it is far smaller than the set of all subgroups of S_n: HARDY [VI.73] and RAMANUJAN [VI.82] showed that the number of partitions of n is about $(1/4n\sqrt{3})\mathrm{e}^{\pi\sqrt{(2n/3)}}$.

Other linearizations are also important. For example, if G acts continuously on a topological space X, we can consider how it acts on homology classes and hence on the HOMOLOGY GROUPS [IV.6 §4] of X.[6] The simplest case of this is the map $z \to \bar{z}$ of the circle S^1. Since this map squares to the identity map, it gives us an action of $\mathbb{Z}/2\mathbb{Z}$ on S^1, which becomes a representation of $\mathbb{Z}/2\mathbb{Z}$ on $H_1(S^1) = \mathbb{R}$ (which represents the identity as multiplication by 1 and the other element of $\mathbb{Z}/2\mathbb{Z}$ as multiplication by -1).

Methods like these have been used to determine the character tables of all finite SIMPLE GROUPS [I.3 §3.3], but they still fall short of a uniform description valid for all groups.

There are many arithmetic properties of the character table that hint at properties of the desired non-Abelian Fourier transform. For example, the size of a conjugacy class divides the order of the group, and in fact the dimension of a representation also divides the order of the group. Pursuing this thought leads to an examination of the values of the characters mod p, relating them to the so-called *p-local subgroups*. These are groups of the form $N(Q)/Q$, where Q is a subgroup of G, the number of elements of Q is a power of p, and $N(Q)$ is the *normalizer* of Q (defined to be the largest subgroup of G that contains Q as a normal subgroup). When the so-called "p-Sylow subgroup" of G is Abelian, beautiful conjectures of Broué give us an essentially complete picture of the representations of G. But in general these questions are at the center of a great deal of contemporary research.

3 Fourier Analysis

We have justified the study of group actions on vector spaces by explaining that the theory of representations has a nice structure that is not present in the theory of group actions on sets. A more historically based account would start by saying that spaces of functions very often come with natural actions of some group G, and many problems of traditional interest can be related to the decomposition of these representations of G.

In this section we will concentrate on the case where G is a compact LIE GROUP [III.48 §1]. We will see that in this case many of the nice features of the representation theory of finite groups persist.

The prototypical example is the space $L^2(S^1)$ of square-integrable functions on the circle S^1. We can think of the circle as the unit circle in $\mathbb{C}$, and thereby identify it with the group of rotations of the circle (since multiplication by $e^{i\theta}$ rotates the circle by θ). This action linearizes to an action on $L^2(S^1)$: if f is a square-integrable function defined on S^1 and w belongs to the circle, then $(w \cdot f)(z)$ is defined to be $f(w^{-1}z)$. That is, $w \cdot f$ does to wz what f does to z.

Classical Fourier analysis expands functions in the space $L^2(S^1)$ in terms of a basis of trigonometric functions: the functions z^n for $n \in \mathbb{Z}$. (These look more "trigonometric" if one writes $e^{i\theta}$ for z and $e^{in\theta}$ for z^n.) If we fix w and write $\phi_n(z) = z^n$, then $(w \cdot \phi_n)(z) = \phi_n(w^{-1}z) = w^{-n}\phi_n(z)$. In particular, $w \cdot \phi_n$ is a multiple of ϕ_n for each w, so the one-dimensional subspace generated by ϕ_n is invariant under the action of S^1. In fact, *every* irreducible representation of S^1 is of this form, as long as we restrict attention to continuous representations.

Now let us consider an innocuous-looking generalization of the above situation: we shall replace 1 by n and try to understand $L^2(S^n)$, the space of complex-valued square-integrable functions on the n-sphere S^n. The n-sphere is acted on by the group of rotations $\mathrm{SO}(n+1)$. As usual, this can be converted into a representation of $\mathrm{SO}(n+1)$ on the space $L^2(S^n)$, which we would like to decompose into irreducible representations; equivalently, we would like to decompose $L^2(S^n)$ into a direct sum of minimal $\mathrm{SO}(n+1)$-invariant subspaces.

This turns out to be possible, and the proof is very similar to the proof for finite groups. In particular, a compact group such as $\mathrm{SO}(n+1)$ has a natural PROBABILITY MEASURE [III.71 §2] on it (called *Haar measure*) in terms of which we can define averages. Roughly speaking, the only difference between the proof for $\mathrm{SO}(n+1)$ and the proof in the finite case is that we have to replace a few sums by integrals.

The general result that one can prove by this method is the following. If G is a compact group that acts continuously on a compact space X (in the sense that each permutation $\phi(g)$ of X is continuous, and also that $\phi(g)$ varies continuously with g), then $L^2(X)$ splits up into an orthogonal direct sum of finite-dimensional minimal G-invariant subspaces; equivalently, the linearized action of G on $L^2(X)$ splits up into an orthogonal direct sum of irreducible representations, all of which are finite dimensional. The problem of finding a

[6] The homology groups discussed in the article just referred to consist of formal sums of homology classes with integer coefficients. Here, where a vector space is required, we are taking real coefficients.

Hilbert space basis of $L^2(X)$ then splits into two subproblems: we must first determine the irreducible representations of G, a problem which is independent of X, and then determine how many times each of these irreducible representations occurs in $L^2(X)$.

When $G = S^1$ (which we identified with $\mathrm{SO}(2)$) and $X = S^1$ as well, we saw that these irreducible representations were one dimensional. Now let us look at the action of the compact group $\mathrm{SO}(3)$ on S^2. It can be shown that the action of G on $L^2(S^2)$ commutes with the *Laplacian*, the differential operator Δ on $L^2(S^2)$ defined by

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

That is, $g(\Delta f) = \Delta(gf)$ for any $g \in G$ and any (sufficiently smooth) function f. In particular, if f is an eigenfunction for the Laplacian (which means that $\Delta f = \lambda f$ for some $\lambda \in \mathbb{C}$), then for each $g \in \mathrm{SO}(3)$ we have

$$\Delta g f = g \Delta f = g \lambda f = \lambda g f,$$

so gf is also an eigenfunction for Δ. Therefore, the space V_λ of all eigenvectors for the Laplacian with eigenvalue λ is G-invariant. In fact, it turns out that if V_λ is nonzero then the action of G on V_λ is an irreducible representation. Furthermore, each irreducible representation of $\mathrm{SO}(3)$ arises exactly once in this way. More precisely, we have a Hilbert space direct sum,

$$L^2(S^2) = \bigoplus_{n \geqslant 0} V_{2n(2n+2)},$$

and each eigenspace $V_{2n(2n+2)}$ has dimension $2n + 1$. Note that this is a case where the set of eigenvalues is *discrete*. (These eigenspaces are discussed further in SPHERICAL HARMONICS [III.87].)

The nice feature that each irreducible representation appears at most once is rather special to the example $L^2(S^n)$. (For an example where this does not happen, recall that with the regular representation $L^2(G)$ of a finite group G each irreducible representation ρ occurs $\dim \rho$ times in $L^2(G)$.) However, other features are more generic: for example, when a compact Lie group acts differentiably on a space X, then the sum of all the G-invariant subspaces of $L^2(X)$ corresponding to a particular representation is always equal to the set of common eigenvectors of some family of commuting differential operators. (In the example above, there was just one operator, the Laplacian.)

Interesting SPECIAL FUNCTIONS [III.85], such as solutions of certain differential equations, often admit representation-theoretic meaning, for example as matrix coefficients. Their properties can then easily be deduced from general results in functional analysis and representation theory rather than from any calculation. Hypergeometric equations, Bessel equations, and many integrable systems arise in this way.

There is more to say about the similarities between the representation theory of compact groups and that of finite groups. Given a compact group G and an irreducible representation ρ of G, we can again take its trace (since it is finite dimensional) and thereby define its character χ_ρ. Just as before, χ_ρ is constant on each conjugacy class. Finally, "the character table is square," in the sense that the characters of the irreducible representations form an orthonormal basis of the Hilbert space of all square-integrable functions that are conjugation invariant in this sense. (Now, though, the "square matrix" is infinite.) When $G = S^1$ this is the Fourier theorem; when G is finite this is the theorem of section 2.

4 Noncompact Groups, Groups in Characteristic p, and Lie Algebras

The "character table is square" theorem focuses our attention on groups with nice conjugacy-class structure. What happens when we take such a group but relax the requirement that it be compact?

A paradigmatic noncompact group is the real numbers $\mathbb{R}$. Like S^1, $\mathbb{R}$ acts on itself in an obvious way (the real number t is associated with the translation $s \mapsto s + t$), so let us linearize that action in the usual way and look for a decomposition of $L^2(\mathbb{R})$ into $\mathbb{R}$-invariant subspaces.

In this situation we have a *continuous family* of irreducible one-dimensional representations: for each real number λ we can define the function χ_λ by $\chi_\lambda(x) = \mathrm{e}^{2\pi \mathrm{i} \lambda x}$. These functions are not square integrable, but despite this difficulty classical Fourier analysis tells us that we can write an L^2-function in terms of them. However, since the Fourier modes now vary in a continuous family, we can no longer decompose a function as a sum: rather we must use an integral. First, we define the Fourier transform $\hat{f}$ of f by the formula $\hat{f}(\lambda) = \int f(x)\mathrm{e}^{2\pi \mathrm{i}\lambda x}\,\mathrm{d}x$. The desired decomposition of f is then $f(x) = \int \hat{f}(\lambda)\mathrm{e}^{-2\pi \mathrm{i}\lambda x}\,\mathrm{d}\lambda$. This, the *Fourier inversion formula*, tells us that f is a weighted integral of the functions χ_λ. We can also think of it as something like a decomposition of $L^2(\mathbb{R})$ as a "direct integral" (rather than direct sum) of the one-dimensional subspaces generated by the functions χ_λ. However, we must treat this picture with due caution since the functions χ_λ do not belong to $L^2(\mathbb{R})$.

This example indicates what we should expect in general. If X is a space with a measure and G acts continuously on it in a way that preserves the measures of subsets of X (as translations did with subsets of $\mathbb{R}$), then the action of G on X gives rise to a measure μ_X defined on the set of all irreducible representations, and $L^2(X)$ can be decomposed as the integral over all irreducible representations with respect to this measure. A theorem that explicitly describes such a decomposition is called a *Plancherel* theorem for X.

For a more complicated but more typical example, let us look at the action of $\mathrm{SL}_2(\mathbb{R})$ (the group of real 2×2 matrices with determinant 1) on $\mathbb{R}^2$ and see how to decompose $L^2(\mathbb{R}^2)$. As we did when we looked at functions defined on S^2, we shall make use of a differential operator. This involves the small technicality that we should look at smooth functions, and we do not ask for them to be defined at the origin. The appropriate differential operator this time turns out to be the Euler vector field $x(\partial/\partial x) + y(\partial/\partial y)$. It is not hard to check that if f satisfies the condition $f(tx, ty) = t^s f(x, y)$ for every x, y, and $t > 0$, then f is an eigenfunction of this operator with eigenvalue s, and indeed all functions in the eigenspace with this eigenvalue, which we shall denote by W_s, are of this form. We can also split W_s up as $W_s^+ \oplus W_s^-$, where W_s^+ and W_s^- consist of the even and odd functions in W_s, respectively.

The easiest way of analyzing the structure of W_s is to compute the action of the LIE ALGEBRA [III.48 §2] $\mathfrak{sl}_2$. For those readers unfamiliar with Lie algebras, we will say only that the Lie algebra of a Lie group G keeps track of the action of elements of G that are "infinitesimally close to the identity," and that in this case the Lie algebra $\mathfrak{sl}_2$ can be identified with the space of 2×2 matrices of trace 0, with $\left(\begin{smallmatrix} a & b \\ c & -a \end{smallmatrix}\right)$ acting as the differential operator $(-ax - by)(\partial/\partial x) + (-cx + ay)(\partial/\partial y)$.

Every element of W_s is a function on $\mathbb{R}^2$. If we restrict these functions to the unit circle, then we obtain a map from W_s to the space of smooth functions defined on S^1, which turns out to be an isomorphism. We already know that this space has a basis of Fourier modes z^m, which we can now think of as $(x + \mathrm{i}y)^m$, defined when $x^2 + y^2 = 1$. There is a unique extension of this from a function defined on S^1 to a function in W_s, namely the function $w_m(x, y) = (x + \mathrm{i}y)^m (x^2 + y^2)^{(s-m)/2}$. One can then check the following actions of simple matrices on these functions (to do so, recall the association of the matrices with differential operators given in the previous paragraph):

$$\begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \cdot w_m = m w_m,$$

$$\begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & -1 \end{pmatrix} \cdot w_m = (m - s) w_{m+2},$$

$$\begin{pmatrix} 1 & -\mathrm{i} \\ -\mathrm{i} & -1 \end{pmatrix} \cdot w_m = (-m - s) w_{m-2}.$$

It follows that if s is not an integer, then from any function w_m in W_s^+ we can produce all the others using the action of $\mathrm{SL}_2(\mathbb{R})$. Therefore, $\mathrm{SL}_2(\mathbb{R})$ acts irreducibly on W_s^+. Similarly, it acts irreducibly on W_s^-. We have therefore encountered a significant difference between this and the finite/compact case: when G is not compact, irreducible representations of G can be infinite dimensional.

Looking more closely at the formulas for W_s when $s \in \mathbb{Z}$, we see more disturbing differences. In order to understand these, let us distinguish carefully between representations that are *reducible* and representations that are *decomposable*. The former are representations that have nontrivial G-invariant subspaces, whereas the latter are representations where one can decompose the space on which G acts into a direct sum of G-invariant subspaces. Decomposable representations are obviously reducible. In the finite/compact case, we used an averaging process to show that reducible representations are decomposable. Now we do not have a natural probability measure to use for the averaging, and it turns out that there can be reducible representations that are not decomposable.

Indeed, if s is a nonnegative integer, then the subspaces W_s^+ and W_s^- give us an example of this phenomenon. They are indecomposable (in fact, this is true even when s is a negative integer not equal to -1) but they contain an invariant subspace of dimension $s + 1$. Thus, we cannot write the representation as a direct sum of irreducible representations. (One can do something a little bit weaker, however: if we quotient out by the $(s + 1)$-dimensional subspace, then the quotient representation can be decomposed.)

It is important to understand that in order to produce these indecomposable but reducible representations we worked not in the space $L^2(\mathbb{R}^2)$ but in the space of smooth functions on $\mathbb{R}^2$ with the origin removed. For instance, the functions w_m above are not square integrable. If we look just at representations of G that act on subspaces of $L^2(X)$, then we *can* split them up into a direct sum of irreducibles: given a G-invariant subspace, its orthogonal complement is also G-invariant.

It might therefore seem best to ignore the other, rather subtle representations and just look at these ones. But it turns out to be easier to study *all* representations and only later ask which ones occur inside $L^2(X)$. For $\mathrm{SL}_2(\mathbb{R})$, the representations we have just constructed (which were subquotients of $W_s^{\pm}$) exhaust all the irreducible representations,[7] and there is a Plancherel formula for $L^2(\mathbb{R}^2)$ that tells us which ones appear in $L^2(\mathbb{R}^2)$ and with what multiplicity:

$$L^2(\mathbb{R}^2) = \int_{-\infty}^{\infty} W_{-1+it}\mathrm{e}^{\mathrm{i}t}\,\mathrm{d}t.$$

To summarize: if G is not compact, then we can no longer take averages over G. This has various consequences:

Representations occur in continuous families. The decomposition of $L^2(X)$ takes the form of a direct integral, not a direct sum.

Representations do not split up into a direct sum of irreducibles. Even when a representation admits a finite composition series, as with the action of $\mathrm{SL}^2(\mathbb{R})$ on $W_s^{\pm}$, it need not split up into a direct sum. So to describe all representations we need to do more than just describe the irreducibles—we also need to describe the glue that holds them together.

So far, the theory of representations of a noncompact group G seems to have *none* of the pleasant features of the compact case. But one thing does survive: there is still an analogue of the theorem that the character table is square. Indeed, we can still define characters in terms of the traces of group elements. But now we must be careful, since the irreducible representation may be on an infinite-dimensional vector space, so that its trace cannot be defined so easily. In fact, characters are not functions on G, but only DISTRIBUTIONS [III.18]. The character of a representation determines the *semisimplification* of a representation ρ: that is, it tells us which irreducible representations are part of ρ, but not how they are glued together.[8]

These phenomena were discovered by Harish-Chandra in the 1950s in an extraordinary series of works that completely described the representation theory of Lie groups such as the ones we have discussed (the precise condition is that they should be real and reductive—a concept that will be explained later in this article) and the generalizations of classical theorems of Fourier analysis to this setting.[9]

Independently and slightly earlier, Brauer had investigated the representation theory of *finite* groups on finite-dimensional vector spaces over fields of characteristic p. Here, too, reducible representations need not decompose as direct sums, though in this case the problem is not lack of compactness (obviously, since everything is finite) but an inability to *average* over the group: we would like to divide by $|G|$, but often this is zero. A simple example that illustrates this is the action of $\mathbb{Z}/p\mathbb{Z}$ on the space $\mathbb{F}_p^2$ that takes x to the 2 × 2 matrix $(\begin{smallmatrix}1 & x\\ 1 & 0\end{smallmatrix})$. This is reducible, since the column vector $(\begin{smallmatrix}1\\ 0\end{smallmatrix})$ is fixed by the action, and therefore generates an invariant subspace. However, if one could decompose the action, then the matrices $(\begin{smallmatrix}1 & x\\ 1 & 0\end{smallmatrix})$ would all be diagonalizable, which they are not.

It is possible for there to be infinitely many indecomposable representations, which again may vary in families. However, as before, there are only finitely many *irreducible* representations, so there is some chance of a "character table is square" theorem in which the rows of the square are parametrized by characters of irreducible representations. Brauer proved just such a theorem, pairing the characters with p*-semisimple* conjugacy classes in G: that is, conjugacy classes of elements whose order is not divisible by p.

We will draw two crude morals from the work of Harish-Chandra and of Brauer. The first is that the category of representations of a group is always a reasonable object, but when the representations are infinite dimensional it requires serious technical work to set it up. Objects in this category do not necessarily decompose as a direct sum of irreducibles (one says that the category is not *semisimple*), and can occur in infinite families, but irreducible objects pair off in some precise way with certain "diagonalizable" conjugacy classes in the group—there is always some kind of analogue of "the character table is square" theorem.

It turns out that when we consider representations in more general contexts—Lie algebras acting on vector spaces, quantum groups, p-adic groups on infinite-dimensional complex or p-adic vector spaces, etc.—these qualitative features stay the same.

[7] To make this precise requires some care about what we mean by "isomorphic." Because many different topological vector spaces can have the same underlying $\mathfrak{sl}_2$-module, the correct notion is of *infinitesimal* equivalence. Pursuing this notion leads to the category of *Harish-Chandra modules*, a category with good finiteness properties.

[8] It is a major theorem of Harish-Chandra that the distribution that defines a character is given by *analytic* functions on a dense subset of the semisimple elements of the group.

[9] The problem of determining the irreducible *unitary* representations for real reductive groups has still not been solved; the most complete results are due to Vogan.

The second moral is that we should always hope for some "non-Abelian Fourier transform": that is, a set that parametrizes irreducible representations and a description of the character values in terms of this set.

In the case of real reductive groups Harish-Chandra's work provides such an answer, generalizing the Weyl character formula for compact groups; for arbitrary groups no such answer is known. For special classes of groups, there are partially successful general principles (the orbit method, Broué's conjecture), of which the deepest are the extraordinary circle of conjectures known as the Langlands program, which we shall discuss later.

5 Interlude: The Philosophical Lessons of "The Character Table Is Square"

Our basic theorem ("the character table is square") tells us to expect that the category of all irreducible representations of G is interesting when the conjugacy-class structure of G is in some way under control. We will finish this essay by explaining a remarkable family of examples of such groups—the rational points of *reductive* algebraic groups—and their conjectured representation theory, which is described by the *Langlands program*.

An *affine algebraic group* is a subgroup of some group GL_n that is defined by polynomial equations in the matrix coefficients. For example, the determinant of a matrix is a polynomial in the matrix coefficients, so the group SL_n, which consists of all matrices in GL_n with determinant 1, is such a group. Another is SO_n, which is the set of matrices with determinant 1 that satisfy the equation $AA^{\mathrm{T}} = I$.

The above notation did not specify what sort of coefficients we were allowing for the matrices. That vagueness was deliberate. Given an algebraic group G and a field k, let us write $G(k)$ for the group where the coefficients are taken to have values in k. For example, $\mathrm{SL}_n(\mathbb{F}_q)$ is the set of $n \times n$ matrices with coefficients in the finite field $\mathbb{F}_q$ and determinant 1. This group is finite, as is $\mathrm{SO}_n(\mathbb{F}_q)$, while $\mathrm{SL}_n(\mathbb{R})$ and $\mathrm{SO}_n(\mathbb{R})$ are Lie groups. Moreover, $\mathrm{SO}_n(\mathbb{R})$ is compact, while $\mathrm{SL}_n(\mathbb{R})$ is not. So among affine algebraic groups over fields one already finds all three types of groups we have discussed: finite groups, compact Lie groups, and noncompact Lie groups.

We can think of $\mathrm{SL}_n(\mathbb{R})$ as the set of matrices in $\mathrm{SL}_n(\mathbb{C})$ that are equal to their complex conjugates. There is another involution on $\mathrm{SL}_n(\mathbb{C})$ that is a sort of "twisted" form of complex conjugation, where we send a matrix A to the complex conjugate of $(A^{-1})^{\mathrm{T}}$. The fixed points of this new involution (that is, the determinant-1 matrices A such that A equals the complex conjugate of $(A^{-1})^{\mathrm{T}}$) form a group called $\mathrm{SU}_n(\mathbb{R})$. This is also called a *real form* of $\mathrm{SL}_n(\mathbb{C})$,[10] and it is compact.

The groups $\mathrm{SL}_n(\mathbb{F}_q)$ and $\mathrm{SO}_n(\mathbb{F}_q)$ are almost simple groups;[11] the classification of finite simple groups tells us, mysteriously, that all but twenty-six of the finite simple groups are of this form. A much, much easier theorem tells us that the *connected compact* groups are also of this form.

Now, given an algebraic group G, we can also consider the instances $G(\mathbb{Q}_p)$, where $\mathbb{Q}_p$ is the field of p-adic numbers, and also $G(\mathbb{Q})$. For that matter, we may consider $G(k)$ for any other field k, such as the FUNCTION FIELD OF AN ALGEBRAIC VARIETY [V.30]. The lesson of section 4 is that we may hope for all of these many groups to have a good representation theory, but that to obtain it there will be serious "analytic" or "arithmetic" difficulties to overcome, which will depend strongly on the properties of the field k.

Lest the reader adopt too optimistic a viewpoint, we point out that not every affine algebraic group has a nice conjugacy-class structure. For example, let V_n be the set of upper triangular matrices in GL_n with 1s along the diagonal, and let k be $\mathbb{F}_q$. For large n, the conjugacy classes in $V_n(\mathbb{F}_q)$ form large and complex families: to parametrize them sensibly one needs more than n parameters (in other words, they belong to families of dimension greater than n, in an appropriate sense), and it is not in fact known how to parametrize them even for a smallish value of n, such as 11. (It is not obvious that this is a "good" question though.)

More generally, solvable groups tend to have horrible conjugacy-class structure, even when the groups themselves are "sensible." So we might expect their representation theory to be similarly horrible. The best we can hope for is a result that describes the entries of the character table *in terms of* this horrible structure—some kind of non-Abelian Fourier integral. For certain p-groups Kirillov found such a result in the 1960s, as

[10] When we say that $\mathrm{SL}_n(\mathbb{R})$ and $\mathrm{SU}_n(\mathbb{R})$ are both "real forms" of $\mathrm{SL}_n(\mathbb{C})$, what is meant more precisely is that in both cases the group can be described as a subgroup of some group of real matrices that consists of all solutions to a set of polynomial equations, and that when the same set of equations is applied instead to the group of *complex* matrices the result is isomorphic to $\mathrm{SL}_n(\mathbb{C})$.

[11] Which is to say that the quotient of these groups by their center is simple.

an example of the "orbit method," but the general result is not yet known.

On the other hand, groups that are similar to connected compact groups do have a nice conjugacy-class structure: in particular, finite simple groups do. An algebraic group is called *reductive* if $G(\mathbb{C})$ has a compact real form. So, for instance, SL_n is reductive by the existence of the real form $\mathrm{SU}_n(\mathbb{R})$. The groups GL_n and SO_n are also reductive, but V_n is not.[12]

Let us examine the conjugacy classes in the group SU_n. Every matrix in $\mathrm{SU}_n(\mathbb{R})$ can be diagonalized, and two conjugate matrices have the same eigenvalues, up to reordering. Conversely, any two matrices in $\mathrm{SU}_n(\mathbb{R})$ with the same eigenvalues are conjugate. Therefore, the conjugacy classes are parametrized by the quotient of the subgroup of all diagonal matrices by the action of S_n that permutes the entries.

This example can be generalized. Any compact connected group has a *maximal torus* T, that is, a maximal subgroup isomorphic to a product of circles. (In the previous example it was the subgroup of diagonal matrices.) Any two maximal tori are conjugate in G, and any conjugacy class in G intersects T in a unique W-orbit on T, where W is the *Weyl group*, the finite group $N(T)/T$ (where $N(T)$ is the normalizer of T).

The description of conjugacy classes in $G(\bar{k})$, for an algebraically closed field $\bar{k}$, is only a little more complicated. Any element $g \in G(\bar{k})$ admits a JORDAN DECOMPOSITION [III.43]: it can be written as $g = su = us$, where s is conjugate to an element of $T(\bar{k})$ and u is unipotent when considered as an element of $\mathrm{GL}_n(\bar{k})$. (A matrix A is *unipotent* if some power of $A - I$ is zero.) Unipotent elements never intersect compact subgroups. When $G = \mathrm{GL}_n$ this is the usual Jordan decomposition; conjugacy classes of unipotent elements are parametrized by partitions of n, which, as we mentioned in section 2, are precisely the conjugacy classes of $W = S_n$. For general reductive groups, unipotent conjugacy classes are again almost the same thing as conjugacy classes in W.[13] In particular, there are finitely many, independent of $\bar{k}$.

Finally, when k is not algebraically closed, one describes conjugacy classes by a kind of Galois descent; for example, in $\mathrm{GL}_n(k)$, semisimple classes are still determined by their characteristic polynomial, but the fact that this polynomial has coefficients in k constrains the possible conjugacy classes.

The point of describing the conjugacy-class structure in such detail is to describe the representation theory in analogous terms. A crude feature of the conjugacy-class structure is the way it decouples the field k from finite combinatorial data that is attached to G but independent of k—things like W, the lattice defining T, roots, and weights.

The "philosophy" suggested by the theorem that the character table is square suggests that the representation theory should also admit such a decoupling: it should be built out of the representation theory of k^*, which is the analogue of the circle, and out of the combinatorial structure of $G(\bar{k})$ (such as the finite groups W). Moreover, representations should have a "Jordan decomposition":[14] the "unipotent" representations should have some kind of combinatorial complexity but little dependence on k, and compact groups should have no unipotent representations.

The Langlands program provides a description along the lines laid out above, but it goes beyond any of the results we have suggested in that it also describes the entries of the character table. Thus, for this class of examples, it gives us (conjecturally) the hoped-for "non-Abelian Fourier transform."

6 Coda: The Langlands Program

And so we conclude by just hinting at statements. If $G(k)$ is a reductive group, we want to describe an appropriate category of representations for $G(k)$, or at least the character table, which we may think of as a "semisimplification" of that category.

Even when k is finite, it is too much to hope that conjugacy classes in $G(k)$ parametrize irreducible representations. But something not so far off is conjectured, as follows.

To a reductive group G over an algebraically closed field, Langlands attaches another reductive group ${}^{\mathrm{L}}G$, the *Langlands dual*, and conjectures that representations of $G(k)$ will be parametrized by conjugacy classes

[12] The miracle, not relevant for this discussion, is that compact connected groups can be easily classified. Each one is essentially a product of circles and non-Abelian simple compact groups. The latter are parametrized by DYNKIN DIAGRAMS [III.48 §3]. They are SU_n, Sp_{2n}, SO_n, and five others, denoted E_6, E_7, E_8, F_4, and G_2. That is it!

[13] They are different, but related. Precisely, they are given by combinatorial data, Lusztig's *two-sided cells* for the corresponding affine Weyl group.

[14] The first such theorems were proved for $\mathrm{GL}_n(\mathbb{F}_q)$ by Green and Steinberg. However, the notion of Jordan decomposition for characters originates with Brauer, in his work on modular representation theory. It is part of his modular analogue of the "character table is square" theorem, which we mentioned in section 3.

in ${}^{\mathrm{L}}G(\mathbb{C})$.[15] However, these are not conjugacy classes of *elements* of ${}^{\mathrm{L}}G(\mathbb{C})$, as before, but of *homomorphisms* from the Galois group of k to ${}^{\mathrm{L}}G$. The Langlands dual was originally defined in a combinatorial manner, but there is now a conceptual definition. A few examples of pairs $(G, {}^{\mathrm{L}}G)$ are $(\mathrm{GL}_n, \mathrm{GL}_n)$, $(\mathrm{SO}_{2n+1}, \mathrm{Sp}_{2n})$, and $(\mathrm{SL}_n, \mathrm{PGL}_n)$.

In this way the Langlands program describes the representation theory as built out of the structure of G and the arithmetic of k.

Although this description indicates the flavor of the conjectures, it is not quite correct as stated. For instance, one has to modify the Galois group[16] in such a way that the correspondence is true for the group $\mathrm{GL}_1(k) = k^*$. When $k = \mathbb{R}$, we get the representation theory of $\mathbb{R}^*$ (or its compact form S^1), which is Fourier analysis; on the other hand, when k is a p-adic local field, the representation theory of k^* is described by local class field theory. We already see an extraordinary aspect of the Langlands program: it precisely unifies and generalizes harmonic analysis and number theory.

The most compelling versions of the Langlands program are "equivalences of derived categories" between the category of representations and certain geometric objects on the spaces of Langlands parameters. These conjectural statements are the hoped-for Fourier transforms.

Though much progress has been made, a large part of the Langlands program remains to be proved. For finite reductive groups, slightly weaker statements have been proved, mostly by Lusztig. As all but twenty-six of the finite simple groups arise from reductive groups, and as the sporadic groups have had their character tables computed individually, this work already determines the character tables of all the finite simple groups.

For groups over $\mathbb{R}$, the work of Harish-Chandra and later authors again confirms the conjectures. But for other fields, only fragmentary theorems have been proved. There is much still to be done.

Further Reading

A nice introductory text on representation theory is Alperin's *Local Representation Theory* (Cambridge University Press, Cambridge, 1993). As for the Langlands program, the 1979 American Mathematical Society volume titled *Automorphic Forms, Representations, and L-functions* (but universally known as "The Corvallis Proceedings") is more advanced, and as good a place to start as any.

IV.10 Geometric and Combinatorial Group Theory

Martin R. Bridson

1 What Are Combinatorial and Geometric Group Theory?

Groups and geometry are ubiquitous in mathematics, groups because the symmetries (or AUTOMORPHISMS [I.3 §4.1]) of any mathematical object in any context form a group and geometry because it allows one to think intuitively about abstract problems and to organize families of objects into spaces from which one may gain some global insight.

The purpose of this article is to introduce the reader to the study of infinite, discrete groups. I shall discuss both the combinatorial approach to the subject that held sway for much of the twentieth century and the more geometric perspective that has led to an enormous flowering of the subject in the last twenty years. I hope to convince the reader that the study of groups is a concern for all of mathematics rather than something that belongs particularly to the domain of algebra.

The principal focus of *geometric group theory* is the interaction of geometry/topology and group theory, through group actions and through suitable translations of geometric concepts into group theory. One wants to develop and exploit this interaction for the benefit of both geometry/topology and group theory. And, in keeping with our assertion that groups are important throughout mathematics, one hopes to illuminate and solve problems from elsewhere in mathematics by encoding them as problems in group theory.

Geometric group theory acquired a distinct identity in the late 1980s but many of its principal ideas have their roots in the end of the nineteenth century. At that time, low-dimensional topology and *combinatorial group theory* emerged entwined. Roughly speaking, combinatorial group theory is the study of groups defined in terms of *presentations*, that is, by means of generators and relations. In order to follow the rest of this introduction the reader must first understand what these terms mean. Since their definitions would require

[15] The $\mathbb{C}$ here is because we are looking at representations on *complex* vector spaces; if we were looking at representations on vector spaces over some field $\mathbb{F}$, we would take ${}^{\mathrm{L}}G(\mathbb{F})$.

[16] The appropriately modified Galois group is called the Weil–Deligne group.