X. Then one can check that $\langle f, U^n f\rangle = \mu(A \cap T^{-n}A)$. It follows that

$$\langle f, A_{N,M}(f)\rangle = \frac{1}{N-M}\sum_{n=M}^{N-1}\mu(A \cap T^{-n}A).$$

If we let $N - M$ tend to infinity, then $A_{N,M}f$ tends to a U-invariant function g. Since g is U-invariant, $\langle f, g\rangle = \langle U^n f, g\rangle$ for every n, and therefore $\langle f, g\rangle = \langle A_{N,M}(f), g\rangle$ for every N and M, and finally $\langle f, g\rangle = \langle g, g\rangle$. By the Cauchy–Schwarz inequality, this is at least $(\int g(x)\,\mathrm{d}\mu)^2 = (\int f(x)\,\mathrm{d}\mu)^2 = \mu(A)^2$. Therefore, we deduce that

$$\lim_{N-M\to\infty}\frac{1}{N-M}\sum_{n=M}^{N-1}\mu(A \cap T^{-n}A) \geqslant (\mu(A))^2.$$

If you choose two "random sets" of measure $\mu(A)$, then their intersection will typically be $(\mu(A))^2$, so the inequality above is saying that the average intersection of A with $T^{-n}A$ is at least as big as the "expected" intersection. This result, due to Khinchin, gives more precise information about the nature of Poincaré recurrence.

When a unitary operator is defined in terms of a measure-preserving transformation as above, it is natural to ask whether the averages converge not just in the sense of the L^2-norm but also in the more classical sense of convergence almost everywhere. (For a related thought in a different context, see CARLESON'S THEOREM [V.5].) The answer is that they do, as was shown by BIRKHOFF [VI.78] soon after he learned of von Neumann's theorem. He proved that for each integrable function f one could find a function f^* such that $f^*(Tx) = f^*(x)$ for almost every x, and such that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}f(T^n x) = f^*(x)$$

for almost every x. Suppose that the transformation T is ergodic, let $A \subset X$ be a set of positive measure, and let $f(x)$ be the characteristic function of A. It follows from Birkhoff's theorem that for almost every $x \in X$ one has

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}f(T^n x) = \frac{\int f\,\mathrm{d}\mu}{\mu(X)} = \frac{\mu(A)}{\mu(X)}.$$

Since the expression

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}f(T^n x)$$

describes the frequency of visits of $T^n x$ to the set A, we see that in an ergodic system the images $x, Tx, T^2x, \dots$ of a typical point $x \in A$ visit A with a frequency that equals the proportion of the space occupied by A.

The ergodic theorems of von Neumann and Birkhoff have been generalized over the years in many different directions. These far-reaching extensions of ergodic theorems, and more generally the *ergodic method*, have found impressive applications in such diverse fields as statistical mechanics, number theory, probability theory, harmonic analysis, and combinatorics.

Further Reading

Furstenberg, H. 1981. *Recurrence in Ergodic Theory and Combinatorial Number Theory*. M. B. Porter Lectures. Princeton, NJ: Princeton University Press.

Krengel, U. 1985. *Ergodic Theorems*, with a supplement by A. Brunel. De Gruyter Studies in Mathematics, volume 6. Berlin: Walter de Gruyter.

Mackey, G. W. 1974. Ergodic theory and its significance for statistical mechanics and probability theory. *Advances in Mathematics* 12:178–268.

The Fermat–Euler Theorem

See MODULAR ARITHMETIC [III.58]

V.10 Fermat's Last Theorem

Many people, even if they are not mathematicians, are aware of the existence of *Pythagorean triples*: that is, triples of positive integers (x, y, z) such that $x^2 + y^2 = z^2$. These give us examples of right-angled triangles with integer side lengths, of which the best known is the "$(3, 4, 5)$ triangle." For any two integers m and n, we have that $(m^2 - n^2)^2 + (2mn)^2 = (m^2 + n^2)^2$, which gives us an infinite supply of Pythagorean triples, and in fact every Pythagorean triple is a multiple of a triple of this form.

FERMAT [VI.12] asked the very natural question of whether similar triples existed for higher powers: that is, could there be a solution in positive integers of the equation $x^n + y^n = z^n$ for some power $n \geqslant 3$? For instance, is it possible to express a cube as a sum of two other cubes? Or rather, Fermat famously claimed that it was not possible, and that he had a proof that space did not permit him to write down. Over the next three and a half centuries, this problem became the most famous unsolved problem in mathematics. Given the amount of effort that went into it, one can be virtually certain that Fermat did not in fact have a proof: the problem appears to be irreducibly difficult, and solvable only by techniques that were developed much later than Fermat.

The fact that Fermat's question was an easy one to think of does not on its own guarantee that it is interesting. Indeed, in 1816 GAUSS [VI.26] wrote in a letter that he found it too isolated a problem to interest him. At the time, that was a reasonable remark: it is often extremely hard to determine whether a given Diophantine equation has a solution, and it is therefore easy to come up with hard problems of a similar nature to Fermat's last theorem. However, Fermat's last theorem has turned out to be exceptional in ways that even Gauss could not have been expected to foresee, and nobody would now describe it as "isolated."

By the time of Gauss's remark, the problem had been solved for $n = 3$ (by EULER [VI.19]) and $n = 4$ (by Fermat; this is the easiest case). The first serious connection between Fermat's last theorem and more general mathematical concerns came with the work of KUMMER [VI.40] in the middle of the nineteenth century. An important observation that had been made by Euler is that it can be fruitful to study Fermat's last theorem in larger RINGS [III.81 §1], since these, if appropriately chosen, allow one to factorize the polynomial $z^n - y^n$. Indeed, if we write $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ for the nth roots of 1, then we can factorize it as

$$(z - y)(z - \zeta y)(z - \zeta^2 y) \cdots (z - \zeta^{n-1} y). \quad (1)$$

Therefore, if $x^n + y^n = z^n$ then we have two rather different-looking factorizations of x^n inside the ring generated by 1 and ζ (namely the factorization in (1) above, and $xxx \cdots x$), and it is reasonable to hope that this information might be exploited. However, there is a serious problem: the ring generated by 1 and ζ does not enjoy the UNIQUE FACTORIZATION PROPERTY [IV.1 §§4–8], so one's sense of being close to a contradiction when faced with these two factorizations is not well-founded. Kummer, in connection with the search for HIGHER RECIPROCITY LAWS [V.28], had met this difficulty and had defined the notion of an IDEAL [III.81 §2]: very roughly, if you enlarge a ring by adding in Kummer's "ideal numbers," then unique factorization is restored. Using these concepts, Kummer was able to prove Fermat's last theorem for every prime number p that was not a factor of the CLASS NUMBER [IV.1 §7] of the corresponding ring. He called such primes *regular*. This connected Fermat's last theorem with ideas that have belonged to the mainstream of ALGEBRAIC NUMBER THEORY [IV.1] ever since. However, it did not solve the problem, since there are infinitely many irregular primes (though this was not known in Kummer's day).

It turned out that more complicated ideas could be used for individual irregular primes, and eventually an algorithm was developed that could check for any given n whether Fermat's last theorem was true for that n. By the late twentieth century, the theorem had been verified for all exponents up to 4 000 000. However, a general proof came from a very different direction.

The story of the eventual proof by Andrew Wiles has been told many times, so we shall be very brief about it here. Wiles did not study Fermat's last theorem directly, but instead solved an important special case of the *Shimura–Taniyama–Weil conjecture*, which connects ELLIPTIC CURVES [III.21] and MODULAR FORMS [III.59]. The first hint that elliptic curves might be relevant came when Yves Hellegouarch noticed that the elliptic curve $y^2 = x(x - a^p)(x - b^p)$ would have rather unusual properties if $a^p + b^p$ was also a pth power. Gerhard Frey realized that such a curve might be so unusual that it would contradict the Shimura–Taniyama–Weil conjecture. Jean-Pierre Serre came up with a precise statement (the "epsilon conjecture") that would imply this, and Ken Ribet proved Serre's conjecture, thus establishing that Fermat's last theorem was a consequence of the Shimura–Taniyama–Weil conjecture. Wiles suddenly became very interested indeed, and after seven years of intensive and almost secret work he announced a solution to a case of the Shimura–Taniyama–Weil conjecture that was sufficient to prove Fermat's last theorem. It then emerged that Wiles's proof contained a serious mistake, but with the help of Richard Taylor he managed to find an alternative and correct argument for that portion of the proof.

The Shimura–Taniyama–Weil conjecture asserts that "all elliptic curves are modular." We finish by giving a rough idea of what this means. (A few more details can be found in ARITHMETIC GEOMETRY [IV.5].) Associated with any elliptic curve E is a sequence of numbers $a_n(E)$, one for each positive integer n. For each prime p, $a_p(E)$ is related to the number of points on the elliptic curve (mod p); it is easy to derive from these values the values of $a_n(E)$ for composite n. Modular forms are HOLOMORPHIC FUNCTIONS [I.3 §5.6] with certain periodicity properties defined on the upper half-plane; associated with each modular form f is a FOURIER SERIES [III.27] that takes the form

$$f(q) = a_1(f)q + a_2(f)q^2 + a_3(f)q^3 + \cdots .$$

Let us call an elliptic curve E *modular* if there is a modular form f such that $a_p(E) = a_p(f)$ for all but finitely many primes p. If you are presented with an elliptic curve, it is not at all clear how to set about finding a modular form associated with it in this way. However, it always seemed to be possible, even if the phenomenon was a mysterious one. For instance, if E is the elliptic curve $y^2 + y = x^3 - x^2 - 10x - 20$, then there is a modular form f such that $a_p(E) = a_p(f)$ for every prime p apart from 11. This modular form is the unique complex function (up to scaling) that satisfies a certain periodicity property with respect to the group $\Gamma_0(11)$, which consists of all matrices $(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix})$ such that a, b, c, and d are integers, c is a multiple of 11, and the DETERMINANT [III.15] $ad - bc$ is 1. It is far from obvious that a definition of this type should have anything to do with elliptic curves.

Wiles proved that all "semistable" elliptic curves are modular, not by showing how to associate a modular form with each such elliptic curve, but by using a subtle counting argument that guaranteed that the modular form had to exist. The full conjecture was proved a few years later, by Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor, which put the icing on the cake of one of the most celebrated mathematical achievements of all time.

V.11 Fixed Point Theorems

1 Introduction

The following is a variant of a well-known mathematical puzzle. A man is on a train from London to Cambridge and has a bottle of water with him. Prove that there is at least one moment on the journey when the volume of air in the bottle, as a fraction of the volume of the bottle itself, is exactly equal to the fraction of his journey that he has completed. (For instance, the bottle might be two fifths full, and therefore three fifths empty, at the precise moment when he is three fifths of the way from London to Cambridge. Note that we do not assume that the bottle is full at the start of the journey or empty at the end.)

The solution, if you have not seen this sort of question before, is surprisingly simple. For each x between 0 and 1 let $f(x)$ be the proportion of air in the bottle when the proportion of the journey that has been completed is x. Then $0 \leqslant f(x) \leqslant 1$ for every x, since the volume of air in the bottle cannot be negative and cannot exceed the volume of the bottle. If we now set $g(x)$ to be $x - f(x)$, then we see that $g(0) \leqslant 0$ and $g(1) \geqslant 0$. Since $g(x)$ varies continuously with x, there must be some moment at which $g(x) = 0$, so that $f(x) = x$, which is what we wanted.

What we have just proved is a slightly disguised form of one of the simplest of all fixed point theorems. We could state it more formally as follows: if f is a continuous function from the closed interval $[0,1]$ to itself, then there must exist an x such that $f(x) = x$. This x we call a *fixed point* of f. (We deduced the result from the *intermediate value theorem*, a basic result in analysis that states that if g is a continuous function from $[0,1]$ to $\mathbb{R}$ such that $g(0) \leqslant 0$ and $g(1) \geqslant 0$, then there must be some x such that $g(x) = 0$.)

In general, a fixed point theorem is a theorem that asserts that a function that satisfies certain conditions must have a fixed point. There are many such theorems, a small sample of which we shall discuss in this article. On the whole, they tend to have a nonconstructive nature: they establish the existence of a fixed point rather than defining one or telling you how to find it. This is part of the reason that they are important, since there are many examples of equations for which one would like to prove that a solution exists even when one cannot solve it explicitly. As we shall see, one way of going about this is to try to rewrite the equation in the form $f(x) = x$ and apply a fixed point theorem.

2 Brouwer's Fixed Point Theorem

The fixed point theorem we have just proved is the one-dimensional version of *Brouwer's fixed point theorem*, which states that if B^n is the unit ball of $\mathbb{R}^n$ (that is, the set of all $(x_1, \dots, x_n)$ such that $x_1^2 + \cdots + x_n^2 \leqslant 1$) and f is a continuous function from B^n to B^n, then f must have a fixed point. The set B^n is an n-dimensional solid sphere, but all that matters is its topological character, so we could take it to be another shape such as an n-dimensional cube or simplex.

In two dimensions this says that a continuous function from the closed unit disk to itself must have a fixed point. In other words, if you had a circular sheet of rubber on a table and you picked it up and put it back down within the circle where it started, having folded it and stretched it as much as you liked, there would always have to be a point that ended up in the same place as before.

To see why this is true, it is helpful to reformulate the statement. Let $D = B^2$ be the closed unit disk. If we had a continuous function f from D to D with no fixed point, then we could define a continuous function g from D to its boundary ∂D as follows: for each x, follow a straight path from $f(x)$ to x and continue on