CHAPTER

1

Number Theory and Proof

Whole numbers are the oldest objects in mathematics, and they still hold some of its deepest secrets.

IN THIS CHAPTER

1.1	A Prime Factory	STEP II 2002
1.2	How Thinly Spread Are the Primes?	STEP III 2016
1.3	The Smallest Counterexample	STEP III 2015
1.4	e Is Irrational, and So Is e^2	STEP III 2024
1.5	Infinite Descent	STEP III 2022
1.6	Sums of Two Squares	STEP II 2023
1.7	When Factorisation Fails	STEP I 2016
1.8	Counting the Rationals	STEP III 2020
1.9	Tangents to a Cubic	STEP I 2009

Why Whole Numbers?

Mathematics is the queen of the sciences, and number theory is the queen of mathematics.

— Carl Friedrich Gauss

Whole numbers are the first mathematical objects anyone meets. You learned to count before you could read. And yet the simplest questions about them can be fiendishly hard. Is every even number greater than 2 the sum of two primes? Nobody knows: Christian Goldbach asked in 1742, computers have checked it beyond 4×10^{18}, and it has still never been proved. Are there infinitely many pairs of primes that differ by 2, like 11 and 13, or 101 and 103? Nobody knows that either, although in 2013 an almost unknown mathematician called Yitang Zhang stunned the world by proving that there are infinitely many pairs of primes differing by at most 70 million, a number since brought down to 246. Pierre de Fermat scribbled in a margin in 1637 that $x^n + y^n = z^n$ has no solutions in positive whole numbers when $n \geq 3$. It took more than 350 years, and some of the most sophisticated mathematics ever created, before Andrew Wiles proved him right in 1994.

This mixture of questions a child can understand and answers that can take centuries is what makes number theory addictive. It is also why number theory has always been the training ground for *proof*. When you cannot check every case, because there are infinitely many whole numbers, you have to find an argument that covers them all at once. Most of the proof techniques you will use for the rest of your mathematical life were first sharpened on questions about whole numbers.

Pure, useless, and running the internet

In 1940 the Cambridge mathematician G. H. Hardy wrote, with some pride, that number theory was the purest of all mathematics, too remote from ordinary life to be put to any practical use, and so too remote to be put to any *harmful* use either. He could hardly have been more wrong. Every time you buy something online, send a secure message or unlock your phone, you are relying on number theory. The RSA system (1977) depends on the fact that multiplying two large primes is easy but recovering them from their product seems to be impossibly hard. Modern systems use the arithmetic of the very curves you will meet at the end of this chapter. The "useless" mathematics of the eighteenth and nineteenth centuries became the security of the twenty-first.

What you will meet at university

If you study mathematics at university, number theory will be one of your first destinations and one of your last.

- **In your first term** you will almost certainly take a course with a name like *Numbers and Sets* or *Foundations*. It will teach you to write proofs properly, and much of its raw material is number theory: induction, the Euclidean algorithm, modular arithmetic, Fermat's Little Theorem, the Chinese Remainder Theorem, and the surprising fact that some infinite sets are bigger than others.
- **In group theory**, usually in the first year, you will discover that many number-theory facts are shadows of a single theorem about symmetry, Lagrange's Theorem.
- **In the second year** a course in *Number Theory* typically covers quadratic reciprocity (Gauss's "golden theorem"), continued fractions and Pell's equation.
- **Later** you can go further: *algebraic number theory*, where new number systems such as the Gaussian integers rescue unique factorisation; *analytic number theory*, which uses calculus and complex analysis to count primes; and the theory of *elliptic curves*, which lies behind the proof of Fermat's Last Theorem and one of the seven million-dollar Millennium Prize Problems.

The questions that run through this chapter

The nine STEP questions in this chapter circle around three big questions. Keep them in mind as you work; we will come back to them at the end.

1. **How many primes are there, and how are they spread out?** Everyone knows there are infinitely many. But how thinly do they thin out, and what kind of argument can tell us?
2. **How can you prove that a number is *not* a fraction?** Showing that a number is rational just needs an example. Showing that it is irrational means ruling out infinitely many possibilities. How is that ever done?
3. **Which equations have whole-number solutions, and what happens when we change the number system?** Sometimes we can show an equation has no solutions at all. Sometimes the best way to understand whole numbers is to step outside them, into the complex numbers or onto a curve.

A chapter about proof

Running underneath those three questions is a fourth theme: *how to prove things*. Each section has been chosen partly because it shows a different method of proof in action, and between them they make a fair toolkit for a first-year undergraduate.

§	Topic	Way of thinking
1.1	Fermat numbers	An infinite family that *needs* infinitely many primes
1.2	Products of primes	Strong induction; information hidden in binomial coefficients
1.3	$\sqrt{2}$ again	"Take the smallest one": the well-ordering principle
1.4	e and e^2	Trapping a whole number strictly between 0 and 1
1.5	Impossible equations	Infinite descent; working modulo n
1.6	Sums of two squares	Changing the setting: complex numbers
1.7	A strange number system	Questioning what you take for granted
1.8	Listing the fractions	Building a bijection; descent in reverse
1.9	Cubes and curves	Geometry solving arithmetic

How to read this chapter

Each section starts with a STEP question exactly as it was set. Try it yourself first, for at least half an hour, before you read on. You will learn far more from a question you have struggled with than from one you have only read about. When you do read on, the *Ask yourself* boxes are invitations to stop and think before the answer arrives, and the *Pause and try* boxes mark the moments where the key idea is within your reach. Then the *What's really going on* part of each section looks past the exam to the mathematics you will meet at university.

The sections increase roughly in sophistication, but they are largely independent. If one doesn't grip you, move on and come back to it later.

1.1 A Prime Factory

The question

The nth *Fermat number*, F_n, is defined by

$$F_n = 2^{2^n} + 1, \qquad n = 0, 1, 2, \ldots,$$

where 2^{2^n} means 2 raised to the power 2^n.

Calculate F_0, F_1, F_2 and F_3. Show that, for $k = 1$, $k = 2$ and $k = 3$,

$$F_0 F_1 \cdots F_{k-1} = F_k - 2. \qquad (*)$$

Prove, by induction or otherwise, that $(*)$ holds for all $k \geq 1$. Deduce that no two Fermat numbers have a common factor greater than 1.

Hence show that there are infinitely many prime numbers.

The hook

You almost certainly already know that there are infinitely many primes. Euclid proved it around 300 BC, and it is one of the first proofs anyone meets. So why would a STEP question ask you to prove it *again*?

Because this proof works in a completely different way. Euclid's argument says: *give me any finite list of primes and I will show you it is incomplete.* This one says: *here is a single, explicit sequence of numbers, written down in advance, and every one of them contains a prime that none of the others contain.* It is a machine for manufacturing new primes, and the reason it works is a beautifully simple identity that the question walks you towards.

Along the way you will meet numbers that fooled Fermat himself.

Before you start

Read the question through once and notice its shape. It has four stages:

1. compute some examples;
2. spot (well, be told) a pattern, and check it on those examples;
3. prove the pattern in general;
4. use the pattern twice: first to deduce something about common factors, then something about primes.

This "examples → conjecture → proof → consequences" structure is how a great deal of mathematics is actually done, and STEP questions often follow it. When you see it, the early parts are usually not just warm-up: they are where you build the intuition you will need later.

Ask yourself The last line asks about *infinitely many primes*. The middle asks about *common factors*. What could possibly connect "no two Fermat numbers share a factor" with "there are infinitely many primes"?

Hold on to that question. We will come back to it.

Working it through

STAGE 1 Getting a feel for the numbers

Be careful with the notation: 2^{2^n} means $2^{(2^n)}$, not $(2^2)^n$. The exponent is itself a power of 2, so these numbers grow *very* fast.

$$F_0 = 2^1 + 1 = 3, \qquad F_2 = 2^4 + 1 = 17,$$
$$F_1 = 2^2 + 1 = 5, \qquad F_3 = 2^8 + 1 = 257.$$

Ask yourself What do you notice about these four numbers?

They are all prime. (Check 257: it is not divisible by $2, 3, 5, 7, 11$ or 13, and $17^2 = 289 > 257$, so we can stop.) The next one, $F_4 = 2^{16} + 1 = 65537$, is also prime. Pierre de Fermat noticed exactly this in 1640 and conjectured that *every* number of this form is prime. Keep that in mind; we will see later whether he was right.

STAGE 2 Checking the identity

Now check $(*)$ for small k:

- $k = 1$: the left-hand side is just $F_0 = 3$, and $F_1 - 2 = 3$. ✓
- $k = 2$: $F_0 F_1 = 3 \times 5 = 15$, and $F_2 - 2 = 15$. ✓
- $k = 3$: $F_0 F_1 F_2 = 15 \times 17 = 255$, and $F_3 - 2 = 255$. ✓

Ask yourself Look at the step from $k = 2$ to $k = 3$. We took 15, multiplied by 17, and got 255. In terms of Fermat numbers, we computed $(F_2 - 2) \times F_2$ and got $F_3 - 2$. Is that a coincidence?

That observation is the whole proof in miniature. Let's see why it should be true.

STAGE 3 Proving the identity

Pause and try Before reading on, try to show directly that

$$(F_k - 2)\,F_k = F_{k+1} - 2.$$

Hint: write everything in terms of powers of 2. What does $F_k - 2$ look like?

We have $F_k - 2 = 2^{2^k} - 1$ and $F_k = 2^{2^k} + 1$, so

$$(F_k - 2)\,F_k = \left(2^{2^k} - 1\right)\left(2^{2^k} + 1\right).$$

Ask yourself This is a product of the form $(a-1)(a+1)$. What is that?

It is a difference of two squares, $a^2 - 1$. With $a = 2^{2^k}$,

$$a^2 = \left(2^{2^k}\right)^2 = 2^{2\cdot 2^k} = 2^{2^{k+1}},$$

so

$$(F_k - 2)\,F_k = 2^{2^{k+1}} - 1 = F_{k+1} - 2. \tag{$\dagger$}$$

That is the engine. Notice what squaring does: it doubles the exponent, and doubling 2^k gives 2^{k+1}. The definition of the Fermat numbers is perfectly tuned to the difference-of-squares factorisation.

Now the induction is almost automatic.

Ask yourself What exactly do we need to show in the inductive step?

We assume $(*)$ holds for some $k \geq 1$,

$$F_0 F_1 \cdots F_{k-1} = F_k - 2,$$

and we want to show it holds for $k+1$:

$$F_0 F_1 \cdots F_{k-1} F_k = F_{k+1} - 2.$$

Multiply both sides of the assumption by F_k and use $(\dagger)$:

$$F_0 F_1 \cdots F_{k-1} F_k = (F_k - 2)\,F_k = F_{k+1} - 2.$$

We have already checked $k = 1$, so by induction $(*)$ holds for every $k \geq 1$. ∎

Ask yourself Could we have avoided induction altogether?

Yes, and it is worth seeing how. Apply $(\dagger)$ repeatedly, starting from $F_0 - 2 = 1$:

$$F_k - 2 = (F_{k-1} - 2)F_{k-1} = (F_{k-2} - 2)F_{k-2}F_{k-1} = \cdots = (F_0 - 2)\,F_0 F_1 \cdots F_{k-1}.$$

Since $F_0 - 2 = 1$, we are done. This "unrolling" is really the same induction written out in full, but some people find it easier to see.

STAGE 4 No common factors

We want to show that if $m \neq n$ then F_m and F_n have no common factor greater than 1.

Ask yourself Suppose some number d divides both F_m and F_n, with $m < n$. Where does F_m appear in $(*)$?

Take $k = n$ in $(*)$:

$$F_0 F_1 \cdots F_m \cdots F_{n-1} = F_n - 2.$$

Since $m < n$, the number F_m is one of the factors on the left. So if d divides F_m, then d divides the whole left-hand side; that is, d divides $F_n - 2$.

Ask yourself Now d divides F_n *and* $F_n - 2$. What else must it divide?

Their difference: d divides $F_n - (F_n - 2) = 2$. So $d = 1$ or $d = 2$.

Ask yourself Can d be 2?

No: every Fermat number is a power of 2 plus 1, so it is odd. Hence $d = 1$, and no two distinct Fermat numbers share a factor greater than 1. ■

Mathematicians say the Fermat numbers are *pairwise coprime*.

STAGE 5 Infinitely many primes

Now we return to the question we asked at the start.

Ask yourself Every whole number greater than 1 has at least one prime factor. Pick one prime factor of each Fermat number. What can you say about the primes you have picked?

Let p_n be any prime that divides F_n. If two of these were equal, say $p_m = p_n$ with $m \neq n$, then that prime would be a common factor of F_m and F_n greater than 1, which we have just shown is impossible. So the primes

$$p_0,\ p_1,\ p_2,\ p_3,\ \ldots$$

are all *different*. There are infinitely many Fermat numbers, so there are infinitely many different primes. ■

Notice that we never needed to know *which* prime divides each Fermat number, or even whether the Fermat numbers are themselves prime. All we needed was that they never share a factor. Each one is forced to "use up" at least one prime that no other can use.

Looking back

It is worth pausing to see what made this work, because the same ideas will come back many times in this book.

- **The examples weren't decoration.** Spotting that $15 \times 17 = 255$ was "$(F_2 - 2) \times F_2 = F_3 - 2$" *was* the key idea. When a question gives you small cases, ask what each step is doing, not just whether the numbers match.
- **A single identity did all the work.** Everything flowed from $(a-1)(a+1) = a^2 - 1$. Hard-looking STEP questions are often driven by one simple piece of algebra.
- **The final argument is about structure, not size.** We proved there are infinitely many primes without finding a single new one. Instead we showed that an infinite family of numbers *needs* infinitely many primes to build them.

What's really going on

Was Fermat right?

No. In 1732 Leonhard Euler showed that

$$F_5 = 2^{32} + 1 = 4\,294\,967\,297 = 641 \times 6\,700\,417.$$

Euler didn't find 641 by luck. He proved that any prime factor of F_n must have a very special form (for $n = 5$, it must be one more than a multiple of 64), which drastically cut down the candidates. You can prove this yourself in Exercise 4.

Since then, every Fermat number anyone has managed to test, from F_5 up to F_{32}, has turned out to be composite. As far as anyone knows, F_0 to F_4 may be the *only* Fermat primes, but nobody has been able to prove it. It is a genuinely open problem.

The delightful twist is that our proof doesn't care. Prime or composite, the Fermat numbers still produce infinitely many primes.

Fermat primes and constructing polygons

Fermat primes turn up somewhere quite unexpected. The ancient Greeks knew how to construct an equilateral triangle and a regular pentagon with ruler and compass, and notice that $3 = F_0$ and $5 = F_1$. For two thousand years nobody could construct any other regular polygon with a prime number of sides. Then in 1796, aged 19, Carl Friedrich Gauss showed how to construct a regular *17-gon*, and $17 = F_2$.

Gauss went further: a regular N-gon can be constructed with ruler and compass exactly when N is a power of 2 multiplied by *distinct Fermat primes*. (Pierre Wantzel completed the proof in 1837.) So which polygons can be constructed depends on how many Fermat primes there are. You will meet the key calculation behind Gauss's 17-gon in Section 6.2 (STEP II 1991, Question 4), and the full story belongs to *Galois theory*, one of the highlights of an undergraduate degree.

One theorem, many proofs

Why would mathematicians bother proving the same theorem more than once? Because different proofs reveal different things. This proof, usually credited to Christian Goldbach in a 1730 letter to Euler, shows that *any* infinite sequence of pairwise coprime whole numbers greater than 1 forces infinitely many primes. Fermat numbers are just one convenient example. (Exercise 3 gives another.)

It also gives more than Euclid's proof does. The primes $p_0, p_1, \ldots, p_{n-1}$ are n *different* primes, each at most F_{n-1}. So there are at least n primes no bigger than $2^{2^{n-1}} + 1$. That is a *quantitative* statement about how the primes are spread out. It is very weak (the truth is much better), but it is a first step towards one of the great results of nineteenth-century mathematics, the *Prime Number Theorem*, which describes exactly how the primes thin out. Section 1.2 takes the next step.

The thinking skill

Building an argument from a structural property

To show that infinitely many objects of type A exist, find infinitely many objects of type B, each of which needs its own object of type A.

Here A = primes and B = Fermat numbers, and "needs its own" means "has a prime factor not shared with any other". The same pattern appears in counting arguments, in linear algebra (infinitely many independent vectors force infinite dimension) and in analysis.

You also practised *induction*, and in particular the habit of asking *"what does the step from k to $k+1$ actually do?"* That question is often more useful than the formal template.

Try this next

1. **Euclid's version.** Write out Euclid's proof that there are infinitely many primes. In what sense is it "by contradiction", and in what sense is the Fermat-number proof "constructive"?

2. **The last digit.** Show that for $n \geq 2$ the Fermat number F_n ends in the digit 7.

3. **Sylvester's sequence.** Define $a_1 = 2$ and $a_{n+1} = a_n^2 - a_n + 1$, so the sequence begins $2, 3, 7, 43, 1807, \ldots$

   (a) Show that $a_{n+1} - 1 = a_1 a_2 \cdots a_n$.

   (b) Deduce that the terms are pairwise coprime, and hence give another proof that there are infinitely many primes.

   (c) HARDER Show that $\dfrac{1}{a_1} + \dfrac{1}{a_2} + \cdots + \dfrac{1}{a_n} = 1 - \dfrac{1}{a_{n+1} - 1}$.

4 Euler's trick. *This needs Fermat's Little Theorem: if p is prime and $p \nmid a$, then $a^{p-1} \equiv 1 \pmod{p}$.* Suppose a prime p divides F_n.

(a) Show that $2^{2^n} \equiv -1 \pmod{p}$, and hence that $2^{2^{n+1}} \equiv 1 \pmod{p}$.

(b) Let d be the smallest positive integer with $2^d \equiv 1 \pmod{p}$. Show that d divides 2^{n+1} but not 2^n, and conclude that $d = 2^{n+1}$.

(c) Using Fermat's Little Theorem, show that d divides $p - 1$. Deduce that $p \equiv 1 \pmod{2^{n+1}}$.

(d) For $n = 5$, this says any prime factor of F_5 has the form $64k + 1$. Check that 641 does. How many primes of this form are smaller than 641?

5 A bound on the primes. Use the argument in *One theorem, many proofs* to show that the nth prime is at most $2^{2^{n-1}} + 1$. How good is this bound when $n = 5$?

Further reading

- **Martin Aigner & Günter M. Ziegler, *Proofs from THE BOOK*** (Springer), Chapter 1, "Six proofs of the infinity of primes". It includes this proof alongside Euclid's and four others, one of which uses topology. One of the best books you can read before university.
- **Harold Davenport, *The Higher Arithmetic*** (Cambridge University Press), Chapters 1–2. A beautifully written introduction to primes and factorisation, ideal for the summer before university.
- **Michal Křížek, Florian Luca & Lawrence Somer, *17 Lectures on Fermat Numbers*** (Springer, 2001). Everything about Fermat numbers, from Euler's factorisation to Gauss's polygons. Dip into the early chapters.
- **Ian Stewart, *Why Beauty Is Truth*** (Basic Books), the chapter on Gauss. The story of the 17-gon told for a general reader.

1.2 How Thinly Spread Are the Primes?

The question

(i) By considering the binomial expansion of $(1+x)^{2m+1}$, prove that

$$\binom{2m+1}{m} < 2^{2m}$$

for any positive integer m.

(ii) For any positive integers r and s with $r < s$, $P_{r,s}$ is defined as follows: $P_{r,s}$ is the product of all the prime numbers greater than r and less than or equal to s, if there are any such prime numbers; if there are no such prime numbers, then $P_{r,s} = 1$.

For example, $P_{3,7} = 35$, $P_{7,10} = 1$ and $P_{14,18} = 17$.

Show that, for any positive integer m, $P_{m+1,2m+1}$ divides $\binom{2m+1}{m}$, and deduce that

$$P_{m+1,2m+1} < 2^{2m}.$$

(iii) Show that, if $P_{1,k} < 4^k$ for $k = 2, 3, \ldots, 2m$, then $P_{1,2m+1} < 4^{2m+1}$.

(iv) Prove that $P_{1,n} < 4^n$ for $n \geq 2$.

The hook

In Section 1.1 we saw that there are infinitely many primes. But "infinitely many" is a blunt statement. The primes certainly get rarer: there are 25 primes below 100 but only 14 between 900 and 1000. *How* rare do they get?

This question proves something that sounds quite modest: if you multiply together all the primes up to n, the answer is less than 4^n. But that innocent inequality is the first real step towards answering "how rare?", and it is the key lemma in one of the most celebrated short proofs in mathematics, by the young Paul Erdős in 1932.

Before you start

The notation $P_{r,s}$ looks fussy, but the idea is simple: multiply together the primes in a range. Make sure you understand the examples. $P_{3,7}$ uses the primes greater than 3 and at most 7, namely 5 and 7, so it is 35.

Ask yourself Part (iv) asks you to prove a statement for *all* $n \geq 2$. Part (iii) assumes the statement for every k up to $2m$. What kind of induction is that?

It is *strong* induction: to prove the case n, you may assume *all* the earlier cases, not just the one immediately before. Keep that in mind; it is why the question is laid out the way it is.

Working it through

STAGE 1 A binomial coefficient is not too big

Ask yourself What do you get if you put $x = 1$ in the binomial expansion of $(1+x)^{2m+1}$?

You get

$$2^{2m+1} = \binom{2m+1}{0} + \binom{2m+1}{1} + \cdots + \binom{2m+1}{2m+1}.$$

All the terms are positive, so each one is less than 2^{2m+1}. That is not quite good enough; we want 2^{2m}, which is half as much.

Ask yourself Pascal's triangle is symmetric. Which other term in the sum equals $\binom{2m+1}{m}$?

By symmetry, $\binom{2m+1}{m} = \binom{2m+1}{m+1}$, and these are two *different* terms of the sum (since $m \neq m+1$). So

$$2\binom{2m+1}{m} = \binom{2m+1}{m} + \binom{2m+1}{m+1} < 2^{2m+1},$$

with strict inequality because the sum also contains other positive terms such as $\binom{2m+1}{0} = 1$. Dividing by 2 gives $\binom{2m+1}{m} < 2^{2m}$. ■

STAGE 2 Which primes divide that binomial coefficient?

Write the binomial coefficient out:

$$\binom{2m+1}{m} = \frac{(2m+1)!}{m!\,(m+1)!} = \frac{(m+2)(m+3)\cdots(2m+1)}{m!}.$$

Ask yourself Let p be a prime with $m+1 < p \leq 2m+1$. Does p divide the numerator? Can it divide the denominator?

The numerator $(m+2)(m+3)\cdots(2m+1)$ contains p itself as one of its factors, so p divides it. The denominator $m!$ is a product of numbers all smaller than p; since p is prime, it cannot divide a product of numbers none of which it divides. So p divides

the numerator but not the denominator, and therefore p divides the whole binomial coefficient (which, remember, is a whole number).

Pause and try We now know that *each* prime in the range divides $\binom{2m+1}{m}$. Why does it follow that their *product* $P_{m+1,2m+1}$ divides it too? (Careful: if 4 and 6 both divide a number, 24 need not.)

The difference is that our primes are *distinct primes*. If distinct primes $p_1, p_2, \ldots$ each divide N, then their product divides N: write $N = p_1 N_1$; then p_2 divides $p_1 N_1$ but not p_1, so it divides N_1; and so on. (This uses the fact that if a prime divides a product then it divides one of the factors. That fact is worth a moment's respect; Section 1.7 shows a world in which its analogue breaks down.)

So $P_{m+1,2m+1}$ divides $\binom{2m+1}{m}$, and since a divisor of a positive number is no bigger than it,

$$P_{m+1,\,2m+1} \le \binom{2m+1}{m} < 2^{2m}. \qquad \blacksquare$$

STAGE 3 Splitting the product

Ask yourself How can $P_{1,2m+1}$, the product of all primes up to $2m+1$, be built from products we already know about?

Split the primes up to $2m+1$ into those up to $m+1$ and those above $m+1$:

$$P_{1,2m+1} = P_{1,m+1} \times P_{m+1,2m+1}.$$

The second factor is less than $2^{2m} = 4^m$ by Stage 2. For the first, we are allowed to assume $P_{1,k} < 4^k$ for every k from 2 to $2m$, and $k = m+1$ is in that range (because $m \ge 1$ gives $2 \le m+1 \le 2m$). So

$$P_{1,2m+1} < 4^{m+1} \times 4^m = 4^{2m+1}. \qquad \blacksquare$$

This is where strong induction earns its keep: to deal with $2m+1$ we needed the case $m+1$, which is roughly *half* as big, not the case just before.

STAGE 4 All n

Ask yourself Part (iii) handles odd numbers. What happens for an even number $n = 2m+2$? Is it prime?

Check the start: $P_{1,2} = 2 < 16$ and $P_{1,3} = 6 < 64$. Now suppose $P_{1,k} < 4^k$ for all k from 2 up to $n-1$, where $n \ge 4$.

- If n is even, then n is not prime (it is bigger than 2), so the primes up to n are the same as the primes up to $n-1$. Hence $P_{1,n} = P_{1,n-1} < 4^{n-1} < 4^n$.
- If n is odd, write $n = 2m+1$ with $m \ge 2$. Then part (iii) gives $P_{1,n} < 4^n$ directly.

By strong induction, $P_{1,n} < 4^n$ for every $n \geq 2$. ■

Looking back

- **Binomial coefficients carry arithmetic information.** $\binom{2m+1}{m}$ is a single number that is small (less than 4^m) but is forced to be divisible by many primes. Squeezing many primes into a small number tells you there can't be too many of them.
- **Strong induction is the natural tool when a problem halves itself.** The case $2m+1$ depended on the case $m+1$.
- **Small cases need checking separately.** The induction only gets going once $n \geq 4$, so $n = 2$ and 3 had to be done by hand.

What's really going on

From 4^n to the Prime Number Theorem

Take logarithms of the result: $\ln P_{1,n} < n \ln 4$, that is,

$$\sum_{p \leq n} \ln p < (\ln 4)\, n \approx 1.39\, n.$$

In the 1850s Pafnuty Chebyshev proved bounds of exactly this kind, above *and below*, showing that $\sum_{p \leq n} \ln p$ grows in proportion to n. Gauss, aged about 15, had already guessed from tables that the number of primes up to n is roughly $n / \ln n$. This became the *Prime Number Theorem*, finally proved in 1896 by Jacques Hadamard and Charles de la Vallée Poussin using complex analysis. In its logarithmic form it says

$$\sum_{p \leq n} \ln p \sim n,$$

so the product of the primes up to n is roughly e^n. Our bound of 4^n is not far off at all: Figure 1.2.1 shows all three.

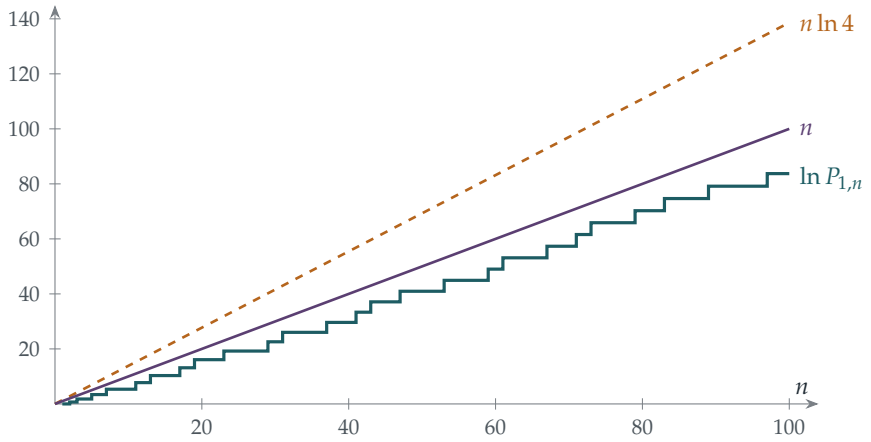

Figure 1.2.1 The logarithm of the product of all primes up to n (teal), jumping by $\ln p$ at each prime p. This section proves it stays below $n \ln 4$ (amber, dashed); the Prime Number Theorem says it grows like n (plum).

Bertrand's postulate

In 1845 Joseph Bertrand conjectured that for every $n \geq 1$ there is always a prime between n and $2n$. Chebyshev proved it in 1852. Eighty years later, Paul Erdős, aged 19, found a beautiful elementary proof. Its key ingredients are this question's lemma and a careful look at which primes divide the *middle* binomial coefficient $\binom{2n}{n}$. Martin Aigner and Günter Ziegler chose it as Chapter 2 of *Proofs from THE BOOK*, a collection of the most elegant proofs in mathematics.

Where this goes next

The precise distribution of the primes is controlled by the zeros of the Riemann zeta function,

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots,$$

a function you will meet in a much humbler guise in Section 2.4, where $\zeta(2) = \pi^2/6$. The *Riemann Hypothesis*, which predicts exactly where those zeros lie, would pin down the error in the Prime Number Theorem as tightly as possible. It is perhaps the most famous unsolved problem in mathematics, and one of the Millennium Prize Problems.

The thinking skill

Strong induction, and extracting information from a clever object

If the case n depends on cases much smaller than $n-1$, assume all earlier cases at once.

The other idea here is subtler and very characteristic of number theory: to show something about *all* primes in a range, find one object (here a binomial coefficient) that every one of those primes must divide, and that you can also show is not too large.

Try this next

1 **Feel the numbers.** Compute $P_{1,n}$ for $n = 10, 20, 30$ and compare it with 4^n and with e^n. (A calculator is fine here.)

2 **The middle binomial coefficient.** Show that $\binom{2m}{m}$ is the largest of the $2m+1$ terms in the expansion of $(1+1)^{2m}$, and deduce that $\binom{2m}{m} \geq \dfrac{4^m}{2m+1}$.

3 **Not too many primes.** Let $\pi(n)$ be the number of primes up to n. Every prime between $\sqrt{n}$ and n is bigger than $\sqrt{n}$, and their product is at most $P_{1,n} < 4^n$. Deduce that

$$\pi(n) - \pi(\sqrt{n}) < \frac{2\ln 4}{\ln n}\, n,$$

and hence that $\pi(n) < 2.8\, n/\ln n + \sqrt{n}$. Compare this with the Prime Number Theorem's $\pi(n) \approx n/\ln n$.

4 **Check Bertrand.** Verify Bertrand's postulate for $n = 1, 2, \ldots, 50$ by finding a prime between n and $2n$ in each case. Can you do it efficiently using a short list of primes, each less than twice the one before?

Further reading

- **Aigner & Ziegler, *Proofs from THE BOOK*,** Chapter 2, "Bertrand's postulate".
- **Marcus du Sautoy, *The Music of the Primes*** (Fourth Estate). A popular history of the Prime Number Theorem and the Riemann Hypothesis.
- **G. J. O. Jameson, *The Prime Number Theorem*** (Cambridge University Press). Rigorous but gentle, and written for undergraduates; the first chapters need only A-level calculus.

1.3 The Smallest Counterexample

The question

(i) In the following argument to show that $\sqrt{2}$ is irrational, give proofs appropriate for steps 3, 5 and 6.

1. Assume that $\sqrt{2}$ is rational.
2. Define the set S to be the set of positive integers with the following property: n is in S if and only if $n\sqrt{2}$ is an integer.
3. Show that the set S contains at least one positive integer.
4. Define the integer k to be the smallest positive integer in S.
5. Show that $(\sqrt{2}-1)k$ is in S.
6. Show that steps 4 and 5 are contradictory and hence that $\sqrt{2}$ is irrational.

(ii) Prove that $2^{1/3}$ is rational if and only if $2^{2/3}$ is rational.

Use an argument similar to that of part (i) to prove that $2^{1/3}$ and $2^{2/3}$ are irrational.

The hook

You have probably seen a proof that $\sqrt{2}$ is irrational: suppose $\sqrt{2} = a/b$ in lowest terms, square, deduce that a is even, then that b is even, contradiction. It is a fine proof. But it relies on a trick about *even numbers* that doesn't obviously generalise. What about $\sqrt[3]{2}$? Or $\sqrt{7}$?

This question offers a different proof, and its central idea is almost absurdly simple: *look at the smallest one*. The same idea will reappear, in disguise, several times in this chapter.

Before you start

This is an unusual STEP question: it gives you the skeleton of a proof and asks you to fill in the missing justifications. That is exactly how undergraduates learn to read mathematics. A textbook proof often says "clearly" or "it follows that", and part of your job is to check that it really does.

Ask yourself Step 4 says "define k to be the smallest positive integer in S". Why are we entitled to do that? Would it be allowed if S were a set of positive *real* numbers?

Any non-empty set of positive integers has a smallest element. This is called the *well-ordering principle*, and it is equivalent to the principle of mathematical induction. It is *false* for sets of positive reals: the set $\{x : x > 0\}$ has no smallest element. So step 4 relies on step 3 (the set must be non-empty) and on S consisting of integers.

Working it through

STAGE 1 Step 3: S is not empty

Ask yourself If $\sqrt{2} = a/b$ for positive integers a and b, can you write down an element of S?

If $\sqrt{2} = a/b$, then $b\sqrt{2} = a$, which is an integer. So b is in S, and S is not empty.

STAGE 2 Step 5: a smaller element

Now k is the smallest element of S, so $k\sqrt{2}$ is an integer; call it j.

Pause and try To show that $(\sqrt{2}-1)k$ belongs to S you must check three things. What are they?

We need $(\sqrt{2}-1)k$ to be (a) an integer, (b) positive, and (c) such that multiplying it by $\sqrt{2}$ gives an integer.

- (a) $(\sqrt{2}-1)k = k\sqrt{2} - k = j - k$, a difference of integers.
- (b) $\sqrt{2} > 1$, so $\sqrt{2} - 1 > 0$ and hence $(\sqrt{2}-1)k > 0$.
- (c) $(\sqrt{2}-1)k \cdot \sqrt{2} = 2k - k\sqrt{2} = 2k - j$, an integer.

So $(\sqrt{2}-1)k \in S$.

STAGE 3 Step 6: the contradiction

Ask yourself How big is $\sqrt{2} - 1$?

Since $1 < \sqrt{2} < 2$, we have $0 < \sqrt{2} - 1 < 1$, and so $(\sqrt{2}-1)k < k$. We have found an element of S smaller than k, the smallest element of S. That is impossible, so our assumption in step 1 was false: $\sqrt{2}$ is irrational. ■

Notice that we never needed "lowest terms" or any facts about even and odd numbers.

The whole proof rests on three facts: $\sqrt{2} \cdot \sqrt{2}$ is an integer, $\sqrt{2}$ lies strictly between two consecutive integers, and integers have a well-ordering.

STAGE 4 Cube roots: linking the two numbers

Write $c = 2^{1/3}$, so that $c^2 = 2^{2/3}$ and $c^3 = 2$.

> ? **Ask yourself** If c is rational, is c^2? And if c^2 is rational, can you get back to c using only multiplication and division?

If c is rational then so is $c^2 = c \times c$. Conversely, if c^2 is rational then so is $2/c^2 = c^3/c^2 = c$. So $2^{1/3}$ is rational if and only if $2^{2/3}$ is. ■

STAGE 5 Adapting the argument

> ? **Ask yourself** In part (i), the key was that $(\sqrt{2})^2 = 2$ is an integer, so multiplying an element of S by $\sqrt{2}$ twice takes you back to integers. For $c = 2^{1/3}$ you would need to multiply *three* times. How should the set S change?

Suppose c is rational; then c^2 is too. Define

$$S = \{n \text{ a positive integer} : nc \text{ and } nc^2 \text{ are both integers}\}.$$

- S **is not empty.** If $c = a/b$ and $c^2 = a'/b'$, then bb' is in S.
- **Let** k **be its smallest element**, and consider $(c-1)k$. We have $1 < c < 2$ (since $1^3 < 2 < 2^3$), so $0 < (c-1)k < k$.
- **It is an integer:** $(c-1)k = kc - k$.
- **Times** c**:** $(c-1)k \cdot c = kc^2 - kc$, an integer.
- **Times** c^2**:** $(c-1)k \cdot c^2 = kc^3 - kc^2 = 2k - kc^2$, an integer.

So $(c-1)k \in S$ and is smaller than k: contradiction. Hence $2^{1/3}$ is irrational, and so, by Stage 4, is $2^{2/3}$. ■

Looking back

- **"Take the smallest one" is a proof technique.** Assume a counterexample exists, take the smallest, and manufacture a smaller one. This is called the *extremal principle*, or a *minimal counterexample* argument.
- **The set** S **did the bookkeeping.** Instead of juggling numerators and denominators, we collected all the "good" multipliers into one set.
- **Generalising meant asking what really made the proof work.** For $\sqrt{2}$ one multiplication by $\sqrt{2}$ was needed; for $\sqrt[3]{2}$ we needed two. Spotting that told us exactly how to redefine S.

What's really going on

A proof you can draw

There is a lovely picture behind part (i), due to Stanley Tennenbaum. If $\sqrt{2} = a/b$ then $a^2 = 2b^2$: a square of side a has the same area as two squares of side b. Place the two b-squares in opposite corners of the a-square, as in Figure 1.3.1. They overlap in a small square, and leave two small squares uncovered. Equal areas force the overlap to equal the two uncovered squares together, giving a *smaller* pair of integers with the same property. Smaller and smaller pairs can't go on for ever, so no such squares exist. The smaller pair is exactly $(2b - a,\ a - b)$, which is our $(\sqrt{2} - 1)k$ step in disguise.

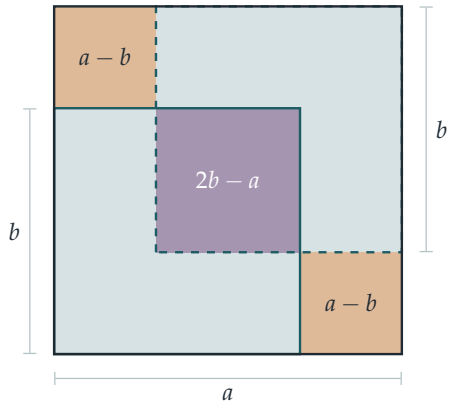

Figure 1.3.1 Tennenbaum's proof. Two squares of side b (teal) sit in opposite corners of a square of side a. If $a^2 = 2b^2$, the area counted twice (the plum overlap) must equal the area missed (the two amber corners), so $(2b - a)^2 = 2(a - b)^2$: a smaller solution.

Well-ordering, induction and descent

The well-ordering principle and the principle of induction are logically equivalent: each can be proved from the other. At university you will see both used as the foundations of the natural numbers. In Section 1.5 we will use the same principle in a different form, Fermat's *method of infinite descent*: there is no infinite strictly decreasing sequence of positive integers.

How far does it go?

The argument proves that $\sqrt{N}$ is irrational whenever N is not a perfect square (see the exercises), and more generally that the nth root of a whole number is either a whole number or irrational. At university this becomes a statement about *algebraic integers*: a rational number that satisfies a monic polynomial equation with integer coefficients, such as $x^3 - 2 = 0$, must be an integer. You will meet the idea again in Section 6.3.

The legend that the Pythagoreans drowned Hippasus for revealing the irrationality of $\sqrt{2}$ is probably untrue. What is true is that the discovery forced Greek mathematicians to separate the study of number from the study of length, a split that took until the nineteenth century, and Dedekind's construction of the real numbers, to heal.

The thinking skill

> **The extremal principle**
>
> *If you want to show that something never happens, suppose it does, choose the smallest (or largest, or simplest) instance, and use it to build an even smaller one.*
>
> The choice of what "smallest" means is the creative step. Here it was the smallest n making $n\sqrt{2}$ an integer.

Try this next

1. **Other square roots.** Adapt the argument to show that $\sqrt{3}$ is irrational. Which number plays the role of $\sqrt{2}-1$?

2. **All non-squares.** Let N be a positive integer that is not a perfect square, and let m be the integer with $m < \sqrt{N} < m+1$. Prove that $\sqrt{N}$ is irrational by considering $(\sqrt{N}-m)k$.

3. **Spot the flaw.** Try to run the argument for $\sqrt{4}$. Exactly which step fails?

4. **Tennenbaum's picture.** Use Figure 1.3.1 to show that if $a^2 = 2b^2$ then $(2b-a)^2 = 2(a-b)^2$. Check this identity algebraically too.

Further reading

- **Daniel Velleman, *How to Prove It*** (Cambridge University Press). A thorough introduction to proof, including induction and well-ordering.
- **R. P. Burn, *A Pathway into Number Theory*** (Cambridge University Press). A book made entirely of carefully sequenced questions; ideal preparation for a first-year course.
- **John Conway & Joseph Shipman, "Extreme proofs I: the irrationality of $\sqrt{2}$"** (*Mathematical Intelligencer*, 2013). Many proofs of one fact, compared with wit.

1.4 e Is Irrational, and So Is e^2

This section uses two questions. The first, from 1997, proves that e is irrational. The second, from 2024, uses a cleverer version of the same idea to prove that e^2 is irrational, which turns out to be genuinely harder.

Warm-up: STEP III 1997, Question 7

For each positive integer n, let

$$a_n = \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} + \cdots,$$
$$b_n = \frac{1}{n+1} + \frac{1}{(n+1)^2} + \frac{1}{(n+1)^3} + \cdots.$$

(i) Evaluate b_n.

(ii) Show that $0 < a_n < 1/n$.

(iii) Deduce that $a_n = n!\,e - [n!\,e]$ (where $[x]$ is the integer part of x).

(iv) Hence show that e is irrational.

Main question: STEP III 2024, Question 7

In this question, you need not consider issues of convergence.

For positive integer n let

$$f(n) = \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} + \cdots$$

and

$$g(n) = \frac{1}{n+1} - \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} - \cdots.$$

(i) Show, by considering a geometric series, that $0 < f(n) < \dfrac{1}{n}$.

(ii) Show, by comparing consecutive terms, that $0 < g(n) < \dfrac{1}{n+1}$.

(iii) Show, for positive integer n, that $(2n)!\,e - f(2n)$ and $\dfrac{(2n)!}{e} + g(2n)$ are both integers.

(iv) Show that if $qe = \dfrac{p}{e}$ for some positive integers p and q, then $qf(2n) + pg(2n)$ is an integer for all positive integers n.

(v) Hence show that the number e^2 is irrational.

The hook

In Section 1.3 we proved that $\sqrt{2}$ is irrational using the fact that $\sqrt{2}$ satisfies a simple equation, $x^2 = 2$. But e doesn't satisfy any such equation that we know of. All we have is its series,

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots.$$

How can a series tell you that a number is not a fraction?

The answer is one of the most useful ideas in number theory: *a whole number cannot lie strictly between* 0 *and* 1. If you can show that some quantity is a whole number, and also that it is strictly between 0 and 1, you have a contradiction. Everything in this section is a variation on that theme.

Before you start

Ask yourself Suppose $e = p/q$. Multiplying e by q gives an integer. What does multiplying by $q!$ do to each term of the series for e?

Multiplying $\dfrac{1}{k!}$ by $q!$ gives an integer whenever $k \leq q$. So $q!\,e$ splits into a whole-number part (the first $q+1$ terms) and a "tail" (everything after). If $e = p/q$ then $q!\,e$ is itself an integer, so the tail must be an integer too. The plan is to show the tail is strictly between 0 and 1.

Working it through

STAGE 1 The warm-up: e is irrational

The series b_n is geometric with first term and ratio both $\frac{1}{n+1}$, so

$$b_n = \frac{1/(n+1)}{1 - 1/(n+1)} = \frac{1}{n}.$$

Ask yourself Compare a_n and b_n term by term. Which is bigger?

The first terms are equal. After that, $(n+1)(n+2) > (n+1)^2$, and in general $(n+$

$1)(n+2)\cdots(n+k) > (n+1)^k$ for $k \geq 2$, so each later term of a_n is smaller than the corresponding term of b_n. Hence $0 < a_n < b_n = 1/n$.

Now look at what a_n really is. Multiply the series for e by $n!$:

$$n!\,e = \underbrace{\frac{n!}{0!} + \frac{n!}{1!} + \cdots + \frac{n!}{n!}}_{\text{an integer}} + \underbrace{\frac{n!}{(n+1)!} + \frac{n!}{(n+2)!} + \cdots}_{=a_n}.$$

Since $0 < a_n < 1$, the integer part of $n!\,e$ is the first bracket and a_n is the fractional part: $a_n = n!\,e - [n!\,e]$.

Pause and try Now finish the proof. Suppose $e = p/q$. Which value of n should you choose?

Take $n = q$. Then $q!\,e = (q-1)!\,p$ is an integer, so its fractional part a_q is 0. But $a_q > 0$. Contradiction: e is irrational. ■

This proof is due to Joseph Fourier, in about 1815.

STAGE 2 Why e^2 is harder

Ask yourself Try the same idea on e^2. Its series is $\sum 2^k/k!$. What goes wrong when you multiply by $n!$ and look at the tail?

The tail is now $\dfrac{2^{n+1}}{n+1} + \dfrac{2^{n+2}}{(n+1)(n+2)} + \cdots$, and the powers of 2 ruin everything: the tail is no longer small. Multiplying by $n!$ doesn't help with the 2^k in the numerators.

The 2024 question's trick is to avoid e^2 altogether. If $e^2 = p/q$, then

$$qe = \frac{p}{e}.$$

Now we have two numbers, e and $1/e$, each with a factorial series and no troublesome powers of 2:

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}, \qquad \frac{1}{e} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}.$$

STAGE 3 The two tails

Part (i) is exactly the warm-up: $0 < f(n) < 1/n$.

For part (ii), $g(n)$ is an *alternating* series whose terms decrease in size.

Ask yourself For an alternating series $t_1 - t_2 + t_3 - \cdots$ with $t_1 > t_2 > t_3 > \cdots > 0$, why is the sum between 0 and t_1?

Group the terms in pairs: $(t_1 - t_2) + (t_3 - t_4) + \cdots$ is a sum of positive brackets, so it is positive. Group them the other way: $t_1 - (t_2 - t_3) - (t_4 - t_5) - \cdots$ is t_1 minus positive brackets, so it is less than t_1. Here $t_1 = \frac{1}{n+1}$, and the terms decrease because each is the previous one divided by something bigger than 1. So $0 < g(n) < \frac{1}{n+1}$.

STAGE 4 Integers again

Multiply e by $(2n)!$ and split off the tail, exactly as in the warm-up:

$$(2n)!\, e = (\text{integer}) + f(2n).$$

So $(2n)!\, e - f(2n)$ is an integer. Now do the same for $1/e$:

$$\frac{(2n)!}{e} = \sum_{k=0}^{2n} (-1)^k \frac{(2n)!}{k!} + \sum_{k=2n+1}^{\infty} (-1)^k \frac{(2n)!}{k!}.$$

Ask yourself What is the sign of the first term of the tail, $k = 2n+1$? How does the tail compare with $g(2n)$?

The first term of the tail has $k = 2n+1$, which is odd, so it is negative. The tail is

$$-\frac{1}{2n+1} + \frac{1}{(2n+1)(2n+2)} - \cdots = -g(2n).$$

So $\dfrac{(2n)!}{e} + g(2n)$ is an integer. ■

This is why the question uses $2n$ rather than n: with an even multiplier, the tail of $1/e$ is *exactly* $-g$, with a predictable sign.

STAGE 5 The contradiction

Suppose $qe = p/e$. Multiply both sides by $(2n)!$ and use Stage 4:

$$q\big(\text{integer} + f(2n)\big) = p\big(\text{integer} - g(2n)\big).$$

Rearranging, $qf(2n) + pg(2n)$ equals a combination of integers, so it is an integer. That is part (iv).

Pause and try Now use the bounds from Stage 3. What happens to $qf(2n) + pg(2n)$ when n is very large?

Both $f(2n)$ and $g(2n)$ are positive, so $qf(2n) + pg(2n) > 0$. And

$$qf(2n) + pg(2n) < \frac{q}{2n} + \frac{p}{2n+1} < \frac{p+q}{2n}.$$

Choose $n > (p+q)/2$. Then $qf(2n) + pg(2n)$ is an integer strictly between 0 and 1. That is impossible, so e^2 is irrational. ■

Looking back

- **The engine is "no integer lies strictly between 0 and 1".** Both proofs build a quantity that must be an integer and then trap it.
- **Factorials make tails small.** Multiplying by $n!$ turned the start of the series into integers and left a tail that shrinks like $1/n$.
- **When a direct attack fails, rearrange.** Rather than attacking $e^2 = p/q$ head-on, we rewrote it as $qe = p/e$, which spread the difficulty across two easier numbers.
- **The details matter.** The choice of $2n$ rather than n fixed the sign of a tail. Many proofs hinge on a small detail like this.

What's really going on

How irrational is e?

Joseph Liouville proved in 1840 that e^2 is irrational, and in fact that it doesn't satisfy any quadratic equation with integer coefficients. In 1873 Charles Hermite went much further: he proved that e is *transcendental*, meaning it is not a root of *any* polynomial with integer coefficients. His proof is a sophisticated version of exactly the argument above: build an integer, show it is strictly between 0 and 1. Nine years later Ferdinand von Lindemann adapted Hermite's method to show that π is transcendental, finally proving that the ancient problem of "squaring the circle" with ruler and compass is impossible.

What we still don't know

For all the power of these methods, embarrassingly simple questions remain open. Is $e + \pi$ irrational? Is $e\pi$? Nobody knows, although at least one of them must be transcendental (exercise 4). Is $\zeta(3) = 1 + \frac{1}{8} + \frac{1}{27} + \cdots$ irrational? Yes: Roger Apéry astonished the mathematical world by proving it in 1978. Is $\zeta(5)$? Nobody knows.

Approximation

Behind these proofs lies a deeper idea: a number is irrational if it can be approximated by fractions *too well*. The partial sums of the series for e are fractions that get very close to e very quickly; a fraction can't be approximated by other fractions that well. This is the starting point of *Diophantine approximation*, a branch of number theory you can explore further in Section 2.9, where a different family of integrals produces even better fractions closing in on e.

The thinking skill

Trapping an integer

To prove a number is irrational, assume it is a fraction, build from it a quantity that must be a whole number, and show that the quantity is strictly between 0 *and* 1.

The art lies in choosing the quantity. Here, multiplying by a factorial made most of a series into integers and left a tail that was small but positive.

Try this next

1. **Using $1/e$ alone.** Prove that e is irrational by supposing $1/e = p/q$ and using only the alternating series for $1/e$.

2. $\sin 1$ **is irrational.** Use $\sin 1 = 1 - \dfrac{1}{3!} + \dfrac{1}{5!} - \cdots$ to prove that $\sin 1$ is irrational. (Multiply by a suitable odd factorial.)

3. **Why not π?** Explain why this method doesn't work directly for π. Then look up Ivan Niven's one-page proof (1947) that π is irrational, which uses an integral instead of a series. It is within reach of a strong A-level student.

4. **One of them is transcendental.** HARDER Show that $e + \pi$ and $e\pi$ are the roots of $x^2 - (e+\pi)x + e\pi = 0$. You may assume that a root of a polynomial whose coefficients are algebraic numbers is itself algebraic, and that π is not algebraic (Lindemann). Deduce that at least one of $e + \pi$ and $e\pi$ is transcendental.

Further reading

- **Aigner & Ziegler, *Proofs from THE BOOK*,** Chapter 8, "Some irrational numbers". Proves that e, e^2, e^4 and π^2 are irrational, using exactly these ideas.
- **Ivan Niven, *Irrational Numbers*** (Mathematical Association of America). A short classic covering everything from $\sqrt{2}$ to transcendence.
- **Julian Havil, *The Irrationals*** (Princeton University Press). A readable history of irrational numbers.

1.5 Infinite Descent

The question

(i) Suppose that there are three non-zero integers a, b and c for which $a^3 + 2b^3 + 4c^3 = 0$.

Explain why there must exist an integer p, with $|p| < |a|$, such that $4p^3 + b^3 + 2c^3 = 0$, and show further that there must exist integers p, q and r, with $|p| < |a|$, $|q| < |b|$ and $|r| < |c|$, such that $p^3 + 2q^3 + 4r^3 = 0$.

Deduce that no such integers a, b and c can exist.

(ii) Prove that there are no non-zero integers a, b and c for which $9a^3 + 10b^3 + 6c^3 = 0$.

(iii) By considering the expression $(3n \pm 1)^2$, prove that, unless an integer is a multiple of three, its square is one more than a multiple of 3. Deduce that the sum of the squares of two integers can only be a multiple of three if each of the integers is a multiple of three.

Hence prove that there are no non-zero integers a, b and c for which $a^2 + b^2 = 3c^2$.

(iv) Prove that there are no non-zero integers a, b and c for which $a^2 + b^2 + c^2 = 4abc$.

The hook

How can you prove that an equation has *no* whole-number solutions? You can't try every possibility; there are infinitely many. Pierre de Fermat, in the 1630s, found a way: show that any solution would produce another solution with *smaller* numbers, and then another, smaller still, and so on for ever. But positive whole numbers can't keep getting smaller for ever. So there was never a solution in the first place.

Fermat called this the *method of infinite descent*, and he was very proud of it. It is the reverse side of the well-ordering principle from Section 1.3, and it is still one of the main tools for proving that equations have no solutions.

Before you start

All four parts have the same shape: assume a solution, shrink it, contradict. The question is *how* to shrink.

Ask yourself In part (i), look at $a^3 + 2b^3 + 4c^3 = 0$. Two of the three terms are obviously even. What does that tell you about the third?

That is the kind of observation that starts every descent: some divisibility is forced on you.

Working it through

STAGE 1 The first descent

Since $a^3 = -2b^3 - 4c^3$, the number a^3 is even, and so a is even (the cube of an odd number is odd). Write $a = 2p$. Then $|p| = |a|/2 < |a|$, and substituting,

$$8p^3 + 2b^3 + 4c^3 = 0 \quad \Longrightarrow \quad 4p^3 + b^3 + 2c^3 = 0.$$

Ask yourself The new equation is $4p^3 + b^3 + 2c^3 = 0$. It has the same coefficients, just rotated. Can you do the same trick again?

Now $b^3 = -4p^3 - 2c^3$ is even, so $b = 2q$ with $|q| < |b|$, and dividing by 2 gives $2p^3 + 4q^3 + c^3 = 0$. Once more, c must be even: $c = 2r$ with $|r| < |c|$, and dividing by 2,

$$p^3 + 2q^3 + 4r^3 = 0.$$

This is the *original* equation, with p, q, r in place of a, b, c. Each is half the size, and none of them is zero, because halving a non-zero number can't give zero.

Pause and try Turn this into a proof that no non-zero solution exists. What, exactly, should you assume is the "smallest"?

Among all solutions in non-zero integers, choose one with $|a|$ as small as possible. (The well-ordering principle lets us do this, if any solutions exist at all.) The argument above produces another non-zero solution with $|p| < |a|$, contradicting minimality. So there are no non-zero solutions. ■

An equivalent way to say it: from one solution we could build an infinite sequence of solutions with $|a| > |p| > \cdots$, an infinite strictly decreasing sequence of positive integers, which is impossible.

STAGE 2 A different prime

Ask yourself In $9a^3 + 10b^3 + 6c^3 = 0$, parity doesn't obviously help. Look at the coefficients 9, 10, 6. Which prime divides two of them?

The prime 3 divides 9 and 6, so 3 divides $10b^3$, and hence 3 divides b. Write $b = 3q$ and

divide by 3:

$$3a^3 + 90q^3 + 2c^3 = 0.$$

Now 3 divides $2c^3$, so $c = 3r$; dividing by 3 again gives $a^3 + 30q^3 + 18r^3 = 0$. So 3 divides a: $a = 3p$, and dividing by 3 once more,

$$9p^3 + 10q^3 + 6r^3 = 0.$$

Exactly the original equation, with smaller non-zero numbers. By the same descent argument, there are no non-zero solutions. ■

STAGE 3 Squares modulo 3

Ask yourself Every integer is of the form $3n$, $3n+1$ or $3n-1$. What are their squares, modulo 3?

$(3n \pm 1)^2 = 9n^2 \pm 6n + 1 = 3(3n^2 \pm 2n) + 1$, one more than a multiple of 3. And $(3n)^2$ is a multiple of 3. So a square leaves remainder 0 or 1 on division by 3, never 2.

Now if $a^2 + b^2$ is a multiple of 3, the possible remainders are $0+0$, $0+1$, $1+0$ and $1+1$, which give 0, 1, 1, 2. Only the first is a multiple of 3, so both a and b must be multiples of 3.

Ask yourself If $a^2 + b^2 = 3c^2$ and $a = 3p$, $b = 3q$, what can you say about c?

We get $9p^2 + 9q^2 = 3c^2$, so $c^2 = 3(p^2 + q^2)$ is a multiple of 3, and hence $c = 3r$. Then $9p^2 + 9q^2 = 27r^2$, so $p^2 + q^2 = 3r^2$: the same equation, smaller numbers. Descent again. ■

STAGE 4 A twist: the equation changes

Ask yourself In $a^2 + b^2 + c^2 = 4abc$ the right-hand side is a multiple of 4. What are the squares modulo 4?

Even squares are multiples of 4 and odd squares are one more than a multiple of 4 ($(2n+1)^2 = 4n^2 + 4n + 1$). So $a^2 + b^2 + c^2$ leaves remainder 0, 1, 2 or 3 on division by 4, according to how many of a, b, c are odd. Since the right-hand side is a multiple of 4, *none* of them can be odd. Write $a = 2p$, $b = 2q$, $c = 2r$:

$$4(p^2 + q^2 + r^2) = 32\,pqr \quad \Longrightarrow \quad p^2 + q^2 + r^2 = 8\,pqr.$$

Ask yourself This is *not* the same equation: the 4 has become an 8. Does the descent still work?

Here is the trick: *prove something stronger*. Consider all equations of the form

$$x^2 + y^2 + z^2 = 2^m\,xyz \qquad \text{with } m \geq 2.$$

For any such m, the right-hand side is a multiple of 4, so exactly the same argument shows x, y, z are all even. Halving them gives a solution of the equation with 2^{m+1}, which is again of this form. So from a non-zero solution of the m equation we get a smaller non-zero solution of the $m+1$ equation. Taking the solution, of *any* equation in the family, with $|x|$ smallest gives a contradiction. In particular, $a^2+b^2+c^2=4abc$ has no non-zero solutions. ■

Sometimes the statement you want to prove is too weak to support its own induction, and the cure is to prove a more general statement that *is* self-supporting. This is a surprisingly common trick at university.

Looking back

- **Descent needs a source of divisibility.** Each time, we found a prime p (here 2 or 3) that was forced to divide all the variables, and divided it out.
- **Working modulo n is the way to find it.** Looking at remainders mod 2, 3 or 4 turned "could anything work?" into a short finite check.
- **If the equation changes, generalise.** Part (iv) needed a whole family of equations, closed under the descent step.

What's really going on

Fermat, descent and his Last Theorem

Fermat used infinite descent to prove that $x^4+y^4=z^2$ has no solutions in positive integers. This is the only case of his famous "Last Theorem" for which he left a proof, and it settles the case $n=4$ of $x^n+y^n=z^n$. Euler later used descent, with a gap that was later filled, for $n=3$. Descent is still the starting point of the modern theory: the proof of the Mordell–Weil theorem about curves like the one in Section 1.9 is a sophisticated descent.

Local obstructions

Every proof in this section worked by looking at remainders modulo some number. If an equation has no solutions modulo n for some n, or no solutions in real numbers, it certainly has no solutions in integers. Such reasons are called *local obstructions*, and at university you will meet the *p-adic numbers*, a way of studying all the remainders modulo $p, p^2, p^3, \ldots$ at once.

A wonderful theorem of Hasse and Minkowski says that for *quadratic* equations like $a^2+b^2=3c^2$, local obstructions are the only obstructions: if there are no "local" reasons to fail, there really is a rational solution. For cubic equations this breaks down spectacularly. Ernst Selmer showed in 1951 that

$$3x^3+4y^3+5z^3=0$$

has solutions modulo every whole number, and in real numbers, but no solution in non-zero integers. Explaining when and why the "local-to-global" principle fails is a major theme of modern number theory.

An equation that *does* have solutions

Part (iv) has a famous cousin: $x^2 + y^2 + z^2 = 3xyz$, the *Markov equation*. It has infinitely many solutions in positive integers, starting with $(1,1,1)$, $(1,1,2)$, $(1,2,5)$, $(1,5,13)$ and $(2,5,29)$, and they are connected to how well irrational numbers can be approximated by fractions. See exercise 4.

The thinking skill

Infinite descent

To show an equation has no solutions, show that any solution would produce a strictly smaller one. Then the smallest solution can't exist.

Pair this with working modulo n to find the divisibility that drives the descent, and be ready to generalise the statement if it isn't preserved by the descent step.

Try this next

1. **A STEP I warm-up.** *(STEP I 2010 Question 8.)* Show that the only integer solution of $a^3 + 3b^3 = 9c^3$ is $a = b = c = 0$. Then, by considering final digits, show the same for $p^4 + 2q^4 = 5r^4$.

2. **A modulus with no descent needed.** Show that $x^2 + y^2 = 4z + 3$ has no integer solutions at all, just by working modulo 4.

3. **Rational points.** Use part (iii) to show that there is no point with *rational* coordinates on the circle $x^2 + y^2 = 3$. (Contrast this with $x^2 + y^2 = 2$, which contains $(1,1)$.)

4. **Markov triples.** Show that if (x, y, z) satisfies $x^2 + y^2 + z^2 = 3xyz$, then so does $(x, y, 3xy - z)$. Starting from $(1,1,1)$, generate the triples listed above. (This "jumping" trick is the opposite of descent.)

Further reading

- **Harold Davenport, *The Higher Arithmetic*,** Chapters 2 and 5.
- **Joseph Silverman, *A Friendly Introduction to Number Theory*** (Pearson). Very readable; includes Fermat's descent proof for $x^4 + y^4 = z^2$.
- **Simon Singh, *Fermat's Last Theorem*** (Fourth Estate). The story, from Fermat's margin

to Wiles's proof, for a general reader.

1.6 Sums of Two Squares

The question

(i) The complex numbers z and w have real and imaginary parts given by $z = a + \mathrm{i}b$ and $w = c + \mathrm{i}d$. Prove that $|zw| = |z||w|$.

(ii) By considering the complex numbers $2 + \mathrm{i}$ and $10 + 11\mathrm{i}$, find positive integers h and k such that $h^2 + k^2 = 5 \times 221$.

(iii) Find positive integers m and n such that $m^2 + n^2 = 8045$.

(iv) You are given that $102^2 + 201^2 = 50805$. Find positive integers p and q such that $p^2 + q^2 = 36 \times 50805$.

(v) Find three distinct pairs of positive integers r and s such that $r^2 + s^2 = 25 \times 1002082$ and $r < s$.

(vi) You are given that $109 \times 9193 = 1002037$. Find positive integers t and u such that $t^2 + u^2 = 9193$.

The hook

Which whole numbers can be written as the sum of two squares? $5 = 1^2 + 2^2$ and $13 = 2^2 + 3^2$ can, but 3, 7 and 21 can't. Here is a strange fact: $65 = 5 \times 13$ can be written as a sum of two squares in *two* different ways, $1^2 + 8^2$ and $4^2 + 7^2$. Why should multiplying two such numbers give another, and why in two ways?

The answer comes from an unexpected direction. The question asks you to leave the whole numbers behind and work with *complex* numbers. In the complex numbers, sums of two squares become simple, and some baffling-looking arithmetic becomes almost mechanical.

Before you start

Ask yourself If $z = a + \mathrm{i}b$ with a and b integers, what is $|z|^2$?

It is $a^2 + b^2$, a sum of two squares. So "N is a sum of two squares" means exactly "$N = |z|^2$ for some complex number z with integer real and imaginary parts". Complex numbers like this are called *Gaussian integers*. Hold on to that translation; the whole

question depends on it.

Working it through

STAGE 1 Moduli multiply

Ask yourself Write out zw in terms of a, b, c, d. What do you need to show about $|zw|^2$?

We have $zw = (ac - bd) + \mathrm{i}(ad + bc)$, so

$$|zw|^2 = (ac - bd)^2 + (ad + bc)^2 = a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2,$$

because the cross terms $-2abcd$ and $+2abcd$ cancel. And this factorises:

$$a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2 = (a^2 + b^2)(c^2 + d^2) = |z|^2|w|^2.$$

Taking square roots, $|zw| = |z||w|$. ■

Now read this as a statement about whole numbers:

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2. \qquad (\star)$$

The product of two sums of two squares is a sum of two squares, and the complex numbers tell you exactly what the two squares are.

STAGE 2 Using it

Part (ii): $5 = |2 + \mathrm{i}|^2$ and $221 = 10^2 + 11^2 = |10 + 11\mathrm{i}|^2$. So multiply:

$$(2 + \mathrm{i})(10 + 11\mathrm{i}) = 20 + 22\mathrm{i} + 10\mathrm{i} - 11 = 9 + 32\mathrm{i},$$

giving $5 \times 221 = 9^2 + 32^2$. (Check: $81 + 1024 = 1105$.)

Ask yourself What happens if you multiply $10 + 11\mathrm{i}$ by $2 - \mathrm{i}$ instead? It has the same modulus as $2 + \mathrm{i}$.

$(2 - \mathrm{i})(10 + 11\mathrm{i}) = 20 + 22\mathrm{i} - 10\mathrm{i} + 11 = 31 + 12\mathrm{i}$, so also $1105 = 31^2 + 12^2$. Choosing $2 + \mathrm{i}$ or its conjugate gives *different* representations. That explains the two ways of writing 65 in the hook (Figure 1.6.1): $65 = |(2 + \mathrm{i})(3 + 2\mathrm{i})|^2 = |4 + 7\mathrm{i}|^2$ and $65 = |(2 - \mathrm{i})(3 + 2\mathrm{i})|^2 = |8 + \mathrm{i}|^2$.

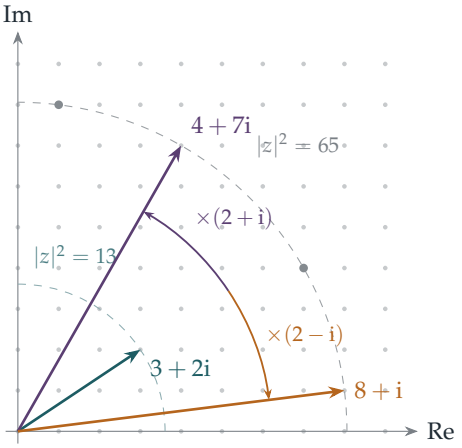

Figure 1.6.1 Multiplying by $2 \pm \mathrm{i}$ scales lengths by $\sqrt{5}$ and rotates by about $26.6°$, anticlockwise for $2+\mathrm{i}$ and clockwise for $2-\mathrm{i}$. Starting from $3+2\mathrm{i}$ on the circle $|z|^2 = 13$, the two choices land on different lattice points of the circle $|z|^2 = 65$. The grey points are their reflections in the line $y = x$.

STAGE 3 Factorise first

Ask yourself 8045 ends in 5. What does that suggest?

$8045 = 5 \times 1609$. We need 1609 as a sum of two squares; trying squares just below 1609 quickly finds $1609 = 40^2 + 3^2$. Then

$$(2+\mathrm{i})(40+3\mathrm{i}) = 77 + 46\mathrm{i}, \qquad \text{so} \qquad 8045 = 77^2 + 46^2,$$

and the conjugate gives $8045 = 83^2 + 34^2$ as well.

Part (iv) is quicker: $36 = 6^2 + 0^2$, and multiplying by 6 just scales: $36 \times 50805 = 612^2 + 1206^2$.

STAGE 4 Three ways

Ask yourself 25 can be written as a sum of two squares in more than one way. How many Gaussian integers have modulus 5, up to signs and swapping?

$25 = 5^2 + 0^2 = 3^2 + 4^2$, corresponding to 5, $3+4\mathrm{i}$ and $3-4\mathrm{i}$. Next, 1002082: since $1001^2 = 1002001$, we have $1002082 = 1001^2 + 9^2$. Multiplying $1001 + 9\mathrm{i}$ by each of the three gives:

$$\begin{aligned} 5\,(1001+9\mathrm{i}) &= 5005 + 45\mathrm{i}, \\ (3+4\mathrm{i})(1001+9\mathrm{i}) &= 2967 + 4031\mathrm{i}, \\ (3-4\mathrm{i})(1001+9\mathrm{i}) &= 3039 - 3977\mathrm{i}. \end{aligned}$$

So the three pairs (r,s) are $(45,5005)$, $(2967,4031)$ and $(3039,3977)$.

STAGE 5 Dividing

Pause and try Part (vi) runs the machine backwards. We know $109 = 10^2 + 3^2$ and $1002037 = 1001^2 + 6^2$. How might you find 9193 as a sum of two squares, without searching?

If multiplying Gaussian integers multiplies moduli, then *dividing* should divide them. Try

$$\frac{1001 + 6\mathrm{i}}{10 + 3\mathrm{i}} = \frac{(1001 + 6\mathrm{i})(10 - 3\mathrm{i})}{109} = \frac{10028 - 2943\mathrm{i}}{109} = 92 - 27\mathrm{i}.$$

It divides exactly, so $9193 = 92^2 + 27^2$. (Check: $8464 + 729 = 9193$.)

Ask yourself What happens if you divide by $10 - 3\mathrm{i}$ instead?

You get $(1001 + 6\mathrm{i})(10 + 3\mathrm{i})/109 = (9992 + 3063\mathrm{i})/109$, which is *not* a Gaussian integer, since $9992/109$ isn't a whole number. Only one of the two "factors of 109" divides $1001 + 6\mathrm{i}$, just as only some primes divide a given whole number. The Gaussian integers have their own primes and their own factorisations.

Looking back

- **Translate, solve, translate back.** "Sum of two squares" became "modulus squared of a Gaussian integer". In that language the problem was easy.
- **Factorise first.** Every part began by breaking a big number into smaller ones whose representations were easy to find.
- **Conjugates give new answers.** The number of representations depends on how many ways there are to choose a factor or its conjugate.

What's really going on

Fermat's two-squares theorem

Which primes are sums of two squares? $2 = 1^2 + 1^2$, 5, 13, 17, 29, 37, 41 are; 3, 7, 11, 19, 23, 31 are not. The pattern: apart from 2, a prime is a sum of two squares exactly when it leaves remainder 1 on division by 4. Fermat stated this in 1640; Euler proved it in 1749. In the language of Gaussian integers it says: a prime $p \equiv 1 \pmod 4$ *splits* as $p = (a + b\mathrm{i})(a - b\mathrm{i})$, while a prime $p \equiv 3 \pmod 4$ stays prime. For example, $5 = (2 + \mathrm{i})(2 - \mathrm{i})$, but 3 cannot be factorised at all.

In 1990 Don Zagier published a proof of the two-squares theorem that is famously one sentence long. It appears, with explanation, in *Proofs from THE BOOK*.

A new number system with its own primes

Gauss realised that the Gaussian integers behave remarkably like the ordinary integers. There is a division algorithm and a Euclidean algorithm, and every Gaussian integer factorises *uniquely* into Gaussian primes (up to multiplying by ± 1 or $\pm \mathrm{i}$). Combined with $(\star)$, this lets you count exactly how many ways a number can be written as a sum of two squares, just from its prime factorisation.

At university this is the gateway to *algebraic number theory*: studying number systems such as $\mathbb{Z}[\mathrm{i}]$ or $\mathbb{Z}[\sqrt{-5}]$ in order to understand ordinary whole numbers. Section 1.7 shows why the subject needed to be invented: not every such system is as well behaved as the Gaussian integers.

Four squares and beyond

Every positive integer is a sum of *four* squares (Lagrange, 1770). The proof relies on a four-square version of $(\star)$, which comes from Hamilton's *quaternions* in the same way that $(\star)$ comes from complex numbers. You will see it in Section 7.8.

The thinking skill

Changing the setting

A problem that is awkward in one number system may be natural in a larger one.

Sums of two squares are awkward in the integers but simply moduli in the complex numbers. Mathematicians constantly enlarge number systems in this way: integers to rationals, rationals to reals, reals to complex numbers. Each enlargement makes some old problems easy.

Try this next

1. **An obstruction.** Show that no number of the form $4k + 3$ is a sum of two squares. (What are the squares modulo 4?)

2. **Two ways.** Write 65 and 85 each as a sum of two squares in two different ways, and explain your answers using Gaussian integers.

3. **Four ways.** $1105 = 5 \times 13 \times 17$. Find four different ways to write 1105 as a sum of two positive squares. Why four?

4. **Gaussian primes.** HARDER Show that if $3 = (a + b\mathrm{i})(c + d\mathrm{i})$ with a, b, c, d integers, then one of the factors has modulus 1. (Take moduli squared.) This shows that 3 is a Gaussian prime.

Further reading

- **John Stillwell, *Elements of Number Theory*** (Springer). Builds up to Gaussian integers and beyond, with history.
- **Aigner & Ziegler, *Proofs from THE BOOK*,** Chapter 4, "Representing numbers as sums of two squares".
- **Joseph Silverman, *A Friendly Introduction to Number Theory*,** chapters on sums of squares and Gaussian integers.

1.7 When Factorisation Fails

The question

The set S consists of all the positive integers that leave a remainder of 1 upon division by 4. The set T consists of all the positive integers that leave a remainder of 3 upon division by 4.

(i) Describe in words the sets $S \cup T$ and $S \cap T$.

(ii) Prove that the product of any two integers in S is also in S. Determine whether the product of any two integers in T is also in T.

(iii) Given an integer in T that is not a prime number, prove that at least one of its prime factors is in T.

(iv) For any set X of positive integers, an integer in X (other than 1) is said to be *X-prime* if it cannot be expressed as the product of two or more integers *all in X* (and all different from 1).

(a) Show that every integer in T is either T-prime or is the product of an odd number of T-prime integers.

(b) Find an example of an integer in S that can be expressed as the product of S-prime integers in two distinct ways. [Note: $s_1 s_2$ and $s_2 s_1$ are not counted as distinct ways of expressing the product of s_1 and s_2.]

The hook

Every whole number greater than 1 can be written as a product of primes in exactly one way: $60 = 2 \times 2 \times 3 \times 5$, and there is no other way to do it. You have probably used this fact so often that it feels too obvious to need proof.

This question builds a small world of numbers that looks perfectly sensible but in which that "obvious" fact is false. It is a STEP I question, so the mathematics is gentle, but the lesson is a deep one. It is exactly the lesson that defeated a famous attempted proof of Fermat's Last Theorem in 1847 and led to the invention of a whole new branch of algebra.

Before you start

Ask yourself List the first few members of S and T. Which members of S are *S-prime*, according to the definition in part (iv)?

$S = \{1, 5, 9, 13, 17, 21, 25, 29, 33, \ldots\}$ and $T = \{3, 7, 11, 15, 19, 23, \ldots\}$. A number in S is S-prime if it can't be written as a product of smaller members of S (other than 1). Notice that 9 is S-prime, even though $9 = 3 \times 3$ in the ordinary integers, because 3 is not in S. "Prime" depends on which numbers you are allowed to use.

Working it through

STAGE 1 Union, intersection and products

$S \cup T$ is the set of all positive odd numbers, since every odd number leaves remainder 1 or 3 on division by 4. $S \cap T$ is empty, because no number leaves two different remainders.

Ask yourself How do you write a general member of S? What is the product of two of them?

Members of S are $4a + 1$ for $a \geq 0$, and

$$(4a + 1)(4b + 1) = 16ab + 4a + 4b + 1 = 4(4ab + a + b) + 1 \in S.$$

So S is closed under multiplication. For T:

$$(4a + 3)(4b + 3) = 16ab + 12a + 12b + 9 = 4(4ab + 3a + 3b + 2) + 1,$$

which is in S, *not* T. For example, $3 \times 7 = 21 \in S$. So T is not closed under multiplication.

STAGE 2 A prime factor in T

Ask yourself Take a number in T that isn't prime. Its prime factors are all odd. What would happen if they were all in S?

The number is odd, so all its prime factors are odd and lie in S or T. If they were *all* in S, then since S is closed under multiplication, their product would be in S. But the product is our number, which is in T. So at least one prime factor is in T. ■

STAGE 3 An odd number of factors

Ask yourself If you multiply together an even number of members of T, where does the product land? What about an odd number?

Each pair of T-numbers multiplies to an S-number, and S-numbers multiply to S-numbers, so an even number of T-factors gives a product in S. An odd number gives (something in S) $\times$ (one element of T), which is in T by the same kind of calculation.

Now take $t \in T$. If t is T-prime, we are done. Otherwise, write t as a product of members of T, each bigger than 1 and so smaller than t. Any of those that are not T-prime can be broken down further, and so on. The factors get smaller each time, so the process stops (well-ordering again), leaving t as a product of T-primes. Since the product is in T, the number of factors must be odd. ■

STAGE 4 Two factorisations

Pause and try Find a number in S with two genuinely different factorisations into S-primes. *Hint:* use numbers like 9, which are S-prime only because their ordinary prime factors are missing from S.

The S-primes begin $5, 9, 13, 17, 21, 29, 33, 37, 41, 49, \ldots$. Now $21 = 3 \times 7$ and $9 = 3 \times 3$ and $49 = 7 \times 7$ are all S-prime, because 3 and 7 are not in S. And

$$441 = 21 \times 21 = 9 \times 49.$$

Two different factorisations of 441 into S-primes (Figure 1.7.1). Unique factorisation fails in S. ■

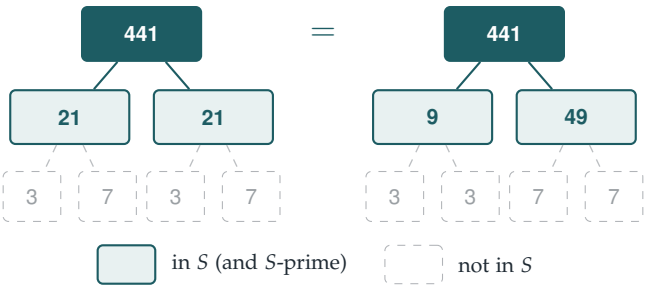

Figure 1.7.1 Two factorisations of 441 into S-primes. In the ordinary integers both trees continue down to $3 \times 3 \times 7 \times 7$, but 3 and 7 are not in S, so inside S the factorisations stop at the second row, and they are different.

Looking back

- **Being "prime" depends on the system.** 9 and 21 are composite in the integers but prime in S, because the numbers you would need to factorise them are missing.
- **Closure matters.** S is closed under multiplication; T is not. That single difference drove parts (iii) and (iv).
- **An obvious fact can be a theorem.** Unique factorisation in the integers needs a proof. This question shows why: in a system that looks very similar, it is false.

What's really going on

What makes the integers special

Why *does* factorisation work in the ordinary integers? The key fact is *Euclid's lemma*: if a prime p divides a product ab, then p divides a or p divides b. It is proved using the Euclidean algorithm, and it is exactly what fails in S. There, the S-prime 9 divides $21 \times 21 = 441$ (because $441 = 9 \times 49$), but 9 doesn't divide 21 in S. In your first year at university you will prove Euclid's lemma carefully and see that unique factorisation follows from it.

The failed proof of Fermat's Last Theorem

On 1 March 1847 Gabriel Lamé announced to the Paris Academy that he had proved Fermat's Last Theorem. His idea was to factorise $x^n + y^n$ using complex roots of unity,

$$x^n + y^n = (x+y)(x+\zeta y)(x+\zeta^2 y)\cdots(x+\zeta^{n-1}y), \qquad \zeta = e^{2\pi \mathrm{i}/n}$$

(for odd n), and then argue about the prime factors of each bracket. Joseph Liouville immediately objected: why should numbers built from ζ factorise uniquely? Weeks later it emerged that Ernst Kummer had already shown that they often don't. The first failure is at $n = 23$.

Kummer rescued what he could by inventing "ideal numbers", extra objects that restore unique factorisation, and proved Fermat's Last Theorem for a large class of exponents. Richard Dedekind turned Kummer's ideal numbers into *ideals*, one of the central concepts of modern algebra.

A smaller example to remember

In the number system $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$,

$$6 = 2 \times 3 = (1+\sqrt{-5})(1-\sqrt{-5}),$$

and all four factors are "prime" in that system (exercise 3). Compare this with the Gaussian integers of Section 1.6, where factorisation *is* unique. Which systems of this kind have unique factorisation is a deep question. Gauss conjectured that, among systems of the form $\mathbb{Z}[\sqrt{-d}]$ and their close relatives, only nine have it. This was finally proved by Heegner, Baker and Stark between 1952 and 1967.

The thinking skill

Questioning what you take for granted

When a fact feels too obvious to prove, ask what it depends on, and look for a setting where it fails.

Building a counterexample, a world where the "obvious" fact is false, is one of the best ways to find out what a proof really needs. Much of undergraduate algebra consists of asking which familiar properties of the integers survive in more general systems.

Try this next

1 **Another example.** Show that 693 has two different factorisations into S-primes. Can you find a general recipe for producing such examples?

2 **Even numbers.** Let $E = \{2, 4, 6, 8, \ldots\}$. Which numbers are E-prime? Show that 60 has more than one factorisation into E-primes.

3 $\mathbb{Z}[\sqrt{-5}]$**.** Define the *norm* $N(a + b\sqrt{-5}) = a^2 + 5b^2$. Show that $N(zw) = N(z)N(w)$. By showing that no element has norm 2 or 3, deduce that 2, 3 and $1 \pm \sqrt{-5}$ cannot be factorised further in $\mathbb{Z}[\sqrt{-5}]$.

4 **Euclid's lemma in action.** HARDER Assuming Euclid's lemma (if a prime divides ab, it divides a or b), prove that every integer greater than 1 has only one factorisation into primes, apart from the order of the factors.

Further reading

- **Ian Stewart & David Tall,** ***Algebraic Number Theory and Fermat's Last Theorem*** (CRC Press). The opening chapters tell the story of Lamé, Kummer and ideals at a level a keen first-year student can follow.
- **John Stillwell,** ***Elements of Number Theory***. Unique factorisation in the integers and in the Gaussian integers, compared side by side.
- **Harold Edwards,** ***Fermat's Last Theorem: A Genetic Introduction*** (Springer). For the ambitious: Kummer's work explained through its history.

1.8 Counting the Rationals

The question

A sequence u_k, for integer $k \geq 1$, is defined as follows.

$$u_1 = 1, \qquad u_{2k} = u_k \text{ for } k \geq 1, \qquad u_{2k+1} = u_k + u_{k+1} \text{ for } k \geq 1.$$

(i) Show that, for every pair of consecutive terms of this sequence, except the first pair, the term with odd subscript is larger than the term with even subscript.

(ii) Suppose that two consecutive terms in this sequence have a common factor greater than one. Show that there are then two consecutive terms earlier in the sequence which have the same common factor. Deduce that any two consecutive terms in this sequence are co-prime (do not have a common factor greater than one).

(iii) Prove that it is not possible for two positive integers to appear consecutively in the same order in two different places in the sequence.

(iv) Suppose that a and b are two co-prime positive integers which do not occur consecutively in the sequence with b following a. If $a > b$, show that $a - b$ and b are two co-prime positive integers which do not occur consecutively in the sequence with b following $a - b$, and whose sum is smaller than $a + b$. Find a similar result for $a < b$.

(v) For each integer $n \geq 1$, define the function f from the positive integers to the positive rational numbers by $f(n) = \dfrac{u_n}{u_{n+1}}$. Show that the range of f is all the positive rational numbers, and that f has an inverse.

The hook

Can you list all the positive fractions, one after another, so that every fraction appears exactly once? It seems impossible: between any two fractions there are infinitely many more, so there is no "next" fraction to write down. Georg Cantor showed in the 1870s that it can nevertheless be done, which means that, in a precise sense, there are no more fractions than there are whole numbers.

Cantor's list was clumsy: it included $\frac{2}{4}$ as well as $\frac{1}{2}$, and the repeats had to be crossed out. This question builds a list with no repeats at all, from a sequence with an absurdly simple rule. It was published by Neil Calkin and Herbert Wilf in 2000, though the

sequence itself was studied by Moritz Stern in 1858.

Before you start

Ask yourself Before proving anything, compute the first fifteen or so terms. What do you notice about the ratios u_n/u_{n+1}?

The sequence begins

$$1,\ 1,\ 2,\ 1,\ 3,\ 2,\ 3,\ 1,\ 4,\ 3,\ 5,\ 2,\ 5,\ 3,\ 4,\ 1,\ 5,\ \ldots$$

and the ratios u_n/u_{n+1} are

$$\tfrac{1}{1},\ \tfrac{1}{2},\ \tfrac{2}{1},\ \tfrac{1}{3},\ \tfrac{3}{2},\ \tfrac{2}{3},\ \tfrac{3}{1},\ \tfrac{1}{4},\ \tfrac{4}{3},\ \tfrac{3}{5},\ \tfrac{5}{2},\ \tfrac{2}{5},\ \tfrac{5}{3},\ \tfrac{3}{4},\ \tfrac{4}{1},\ \ldots$$

Every fraction appears in lowest terms, and none is repeated. The question asks you to prove that this goes on for ever, and that every positive fraction eventually turns up.

Working it through

STAGE 1 Odd beats even

Consecutive terms are either (u_{2k}, u_{2k+1}) or (u_{2k+1}, u_{2k+2}).

Ask yourself Write each of these pairs in terms of u_k and u_{k+1}. Which term is bigger?

$(u_{2k}, u_{2k+1}) = (u_k,\ u_k + u_{k+1})$ and $(u_{2k+1}, u_{2k+2}) = (u_k + u_{k+1},\ u_{k+1})$. Since every term is a positive integer (an easy induction), $u_k + u_{k+1}$ is bigger than both u_k and u_{k+1}. The only exception is the first pair, $(u_1, u_2) = (1, 1)$, which doesn't fit either pattern because u_1 is defined directly. ■

STAGE 2 Consecutive terms are coprime

Ask yourself Suppose $d > 1$ divides both u_{2k} and u_{2k+1}, that is, both u_k and $u_k + u_{k+1}$. What else does d divide?

d divides $(u_k + u_{k+1}) - u_k = u_{k+1}$, so d divides the consecutive pair u_k, u_{k+1}, which is earlier in the sequence. The other case is identical: if d divides $u_k + u_{k+1}$ and u_{k+1}, it divides u_k.

So a common factor of a consecutive pair is passed back to an earlier pair, then an earlier one, and so on. Positions can't decrease for ever (well-ordering yet again), so eventually d would have to divide the first pair, $u_1 = u_2 = 1$. That is impossible for $d > 1$. So every consecutive pair is coprime. ■

STAGE 3 **No pair appears twice**

Pause and try Stage 1 tells you something useful: if you see a consecutive pair (a, b), you can tell whether a sits at an odd or an even position. How? And can you then recover the pair that came before it, in the sense of Stage 2?

If $a < b$, then (by Stage 1) b is at the odd position, so the pair is $(u_{2k}, u_{2k+1}) = (u_k, u_k + u_{k+1})$ for some k, and the "parent" pair (u_k, u_{k+1}) is $(a,\ b-a)$. If $a > b$, the pair is (u_{2k+1}, u_{2k+2}) and its parent is $(a-b,\ b)$. If $a = b$, Stage 1 shows the pair must be the very first pair, $(1, 1)$.

So each consecutive pair other than the first determines its parent pair, *and the position of the parent* (you can see this structure as a tree in Figure 1.8.1): the parent is at position k when the child is at $2k$ or $2k+1$. Now suppose (a, b) appeared at two different positions. Their parents would be the same pair at two positions, and so would *their* parents, and so on. Following this back, we would eventually find $(1, 1)$ at a position other than 1, which Stage 1 rules out. ■

STAGE 4 **Every coprime pair appears**

Ask yourself Suppose some coprime pair (a, b) never appears consecutively. Which such pair should you look at?

Take the "missing" coprime pair with the smallest possible value of $a + b$: a minimal counterexample. It can't be $(1, 1)$, which does appear, so $a \neq b$ (coprime and equal forces $a = b = 1$).

- If $a > b$: the pair $(a - b, b)$ is coprime (any common factor of $a - b$ and b divides a), and its sum a is smaller than $a + b$. By minimality it does appear, say at positions $k, k+1$. But then $u_{2k+1} = u_k + u_{k+1} = a$ and $u_{2k+2} = u_{k+1} = b$, so (a, b) appears after all.
- If $a < b$: similarly $(a, b - a)$ appears at some $k, k+1$, and then $(u_{2k}, u_{2k+1}) = (a, b)$.

Either way, contradiction. So every coprime pair of positive integers appears consecutively. ■

(The question phrases part (iv) as a descent: a missing pair produces a smaller missing pair. That is the same argument, read the other way.)

STAGE 5 **A list of all fractions**

Every positive rational number can be written uniquely in lowest terms as a/b with a, b coprime. By Stage 4, (a, b) appears as some consecutive pair (u_n, u_{n+1}), so $f(n) = a/b$: the range of f is all positive rationals. By Stage 3 it appears only once, and by Stage 2 every value $f(n)$ is already in lowest terms, so two different n can't give the same fraction. Hence f is a bijection from the positive integers to the positive rationals, and has an inverse. ■

Looking back

- **Descent and its reverse.** Common factors and repeated pairs were chased *back* to the start of the sequence. Missing pairs were handled by a minimal counterexample, building *forward*.
- **A bijection needs two halves.** "Every fraction appears" (onto) and "no fraction appears twice" (one-to-one) are separate claims, proved separately.
- **Compute first.** Writing out fifteen terms made the claims believable and suggested the parent–child structure that drove the proofs.

What's really going on

The Calkin–Wilf tree

Arrange the positive fractions in a tree: put $\frac{1}{1}$ at the top, and give each fraction $\frac{a}{b}$ two children, $\frac{a}{a+b}$ on the left and $\frac{a+b}{b}$ on the right. Reading the tree row by row, left to right, gives exactly the list $f(1), f(2), f(3), \ldots$ from this question (Figure 1.8.1). Stage 3 says every fraction has a unique parent, and Stage 4 says every fraction is in the tree. In 2003 Moshe Newman found that you don't even need the tree: each fraction in the list can be computed from the one before by

$$x_{n+1} = \frac{1}{2\lfloor x_n \rfloor - x_n + 1}.$$

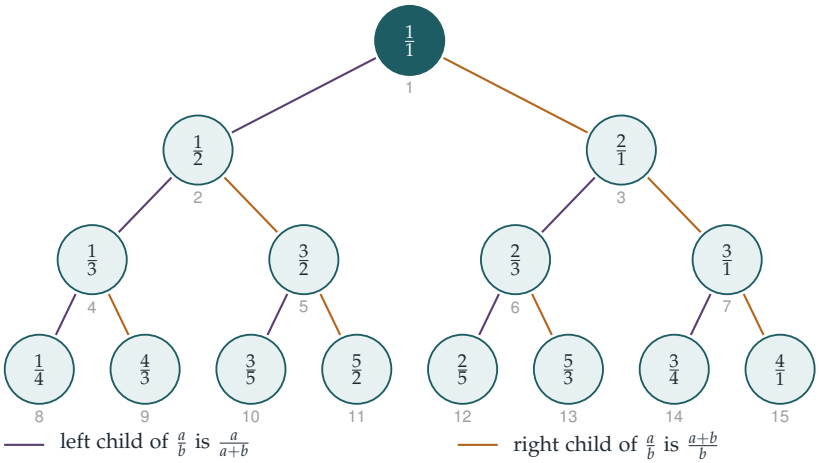

Figure 1.8.1 The first four rows of the Calkin–Wilf tree. Reading row by row, left to right, gives $f(1), f(2), f(3), \ldots$; the small grey numbers are the positions n. Every positive fraction appears exactly once, in lowest terms.

Different sizes of infinity

A set that can be listed like this is called *countable*. The fractions are countable, which is surprising, since they seem so much more numerous than the whole numbers. In your first term at university you will also meet Cantor's *diagonal argument*, which shows that the *real* numbers are not countable: any list of real numbers must miss some. So there really are different sizes of infinity, and the real numbers form a strictly bigger infinity than the rationals. This is one of the most startling results you will meet in a first-year course.

Stern's sequence

The sequence u_n is *Stern's diatomic sequence*, and it has many other lives. For example, u_{n+1} counts the number of ways to write n as a sum of powers of 2 with each power used at most twice. It is closely related to the *Stern–Brocot tree* and to continued fractions, which give the best rational approximations to irrational numbers.

The thinking skill

Building a bijection

To show two sets are the same size, match them up exactly: prove that every element of the target is hit (onto), and that nothing is hit twice (one-to-one).

Bijections are the foundation of counting, finite and infinite, and you will use them constantly at university. Here both halves were proved by chasing a structure back to its starting point.

Try this next

1. **Find the fraction.** For which n is $f(n) = \frac{3}{5}$? And $f(n) = \frac{5}{3}$? Use the parent rule to work backwards to $\frac{1}{1}$.

2. **Newman's formula.** Check that Newman's formula produces the first ten fractions of the list, starting from $x_1 = 1$.

3. **Cantor's diagonal argument.** Suppose you had a list of all real numbers between 0 and 1, written as decimals. Construct a decimal that differs from the nth number in the list in its nth digit. Why does this show the list is incomplete? (What care is needed with decimals like $0.0999\ldots = 0.1$?)

4. **Countable unions.** HARDER Show that the set of all pairs (m, n) of positive integers is countable by giving an explicit bijection with the positive integers. (Try $2^{m-1}(2n-1)$.)

Further reading

- **Neil Calkin & Herbert Wilf, "Recounting the rationals"** (*American Mathematical Monthly*, 2000). Four pages, entirely readable.
- **Aigner & Ziegler, *Proofs from THE BOOK***, Chapter 17, "Sets, functions, and the continuum hypothesis", which includes the Calkin–Wilf tree.
- **Ian Stewart & David Tall, *The Foundations of Mathematics*** (Oxford University Press). Countability, Cantor and the real numbers, as in a first-year course.

1.9 Tangents to a Cubic

The question

A curve has the equation

$$y^3 = x^3 + a^3 + b^3,$$

where a and b are positive constants. Show that the tangent to the curve at the point $(-a, b)$ is

$$b^2y - a^2x = a^3 + b^3.$$

In the case $a = 1$ and $b = 2$, show that the x-coordinates of the points where the tangent meets the curve satisfy

$$7x^3 - 3x^2 - 27x - 17 = 0.$$

Hence find positive integers p, q, r and s such that

$$p^3 = q^3 + r^3 + s^3.$$

The hook

This looks like a routine coordinate-geometry question: differentiate, find a tangent, solve a cubic. Then the last line asks for a cube that is the sum of three other cubes, and the geometry produces one as if by magic.

The trick, drawing a line through a known point on a curve to find a new point, goes back to Diophantus of Alexandria in the third century. In modern hands it became the theory of *elliptic curves*, which sits at the centre of present-day number theory, underlies the proof of Fermat's Last Theorem, and secures a large share of the world's internet traffic.

Before you start

Ask yourself The point $(-a, b)$ is on the curve. Why? And what does that have to do with sums of cubes?

Substitute: $(-a)^3 + a^3 + b^3 = b^3$, so yes. Points on $y^3 = x^3 + a^3 + b^3$ with *rational* coordinates are exactly what we want: if $x = q/s'$ and $y = p/s'$ then clearing denominators gives a relation between cubes. The plan is to start from one easy point and use the geometry to find a less obvious one.

Working it through

STAGE 1 The tangent

Differentiate $y^3 = x^3 + a^3 + b^3$ implicitly: $3y^2\dfrac{dy}{dx} = 3x^2$, so $\dfrac{dy}{dx} = \dfrac{x^2}{y^2}$. At $(-a, b)$ the gradient is a^2/b^2, so the tangent is

$$y - b = \frac{a^2}{b^2}(x + a) \quad\Longrightarrow\quad b^2y - b^3 = a^2x + a^3 \quad\Longrightarrow\quad b^2y - a^2x = a^3 + b^3. \qquad \blacksquare$$

STAGE 2 Where the tangent meets the curve again

With $a = 1$, $b = 2$ the curve is $y^3 = x^3 + 9$ and the tangent at $(-1, 2)$ is $4y - x = 9$, that is, $y = \dfrac{x+9}{4}$. Substitute into the curve:

$$\frac{(x+9)^3}{64} = x^3 + 9 \quad\Longrightarrow\quad x^3 + 27x^2 + 243x + 729 = 64x^3 + 576,$$

which rearranges to $63x^3 - 27x^2 - 243x - 153 = 0$, and dividing by 9 gives $7x^3 - 3x^2 - 27x - 17 = 0$. $\blacksquare$

STAGE 3 Solving the cubic

Ask yourself You already know one root of this cubic without doing any algebra. Which one, and why should it appear *twice*?

The tangent touches the curve at $x = -1$ (Figure 1.9.1), and "touching" means a repeated root, just as a tangent to a parabola meets it at a double root of a quadratic. So $(x+1)^2$ should be a factor. Check: $7(-1)^3 - 3(1) - 27(-1) - 17 = -7 - 3 + 27 - 17 = 0$, and the derivative $21x^2 - 6x - 27$ also vanishes at $x = -1$. Dividing,

$$7x^3 - 3x^2 - 27x - 17 = (x+1)^2(7x - 17).$$

Pause and try So the tangent meets the curve again where $x = \frac{17}{7}$. Find y, and turn the point into a relation between cubes.

$y = \dfrac{17/7 + 9}{4} = \dfrac{80/7}{4} = \dfrac{20}{7}$. The point $\left(\frac{17}{7}, \frac{20}{7}\right)$ lies on $y^3 = x^3 + 9$, so

$$\frac{20^3}{7^3} = \frac{17^3}{7^3} + 9 \quad\Longrightarrow\quad 20^3 = 17^3 + 9 \times 7^3.$$

Ask yourself We want 9×7^3 to be a sum of two cubes. Remember where the 9 came from.

$9 = a^3 + b^3 = 1^3 + 2^3$, so $9 \times 7^3 = 7^3 + 14^3$. Therefore

$$20^3 = 17^3 + 14^3 + 7^3.$$

Check: $8000 = 4913 + 2744 + 343$. ■

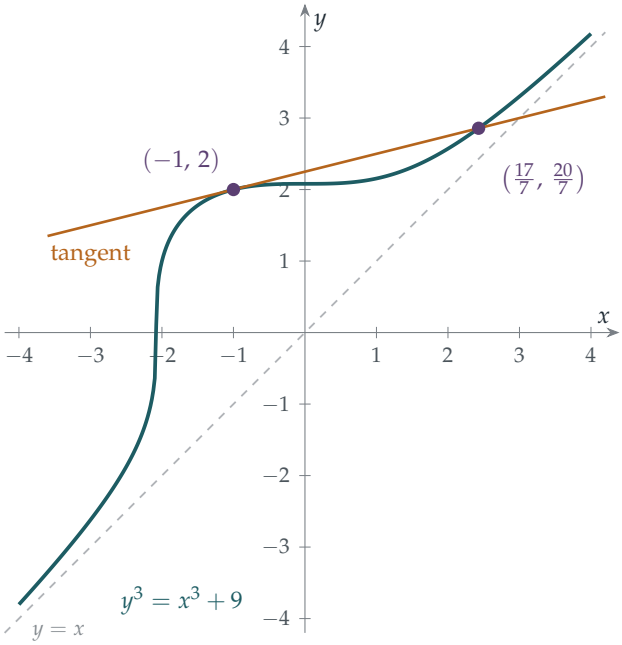

Figure 1.9.1 The curve $y^3 = x^3 + 9$ (teal) approaches the line $y = x$ far from the origin. The tangent at the rational point $(-1, 2)$ (amber) touches the curve there, which counts as a double intersection, and crosses it once more at another rational point, $\left(\frac{17}{7}, \frac{20}{7}\right)$.

Looking back

- **Tangency means a double root.** That turned a cubic with no obvious roots into one we could factorise immediately.
- **The new root is rational for free.** Once two of the three roots are rational (here both -1), the third must be too, because the sum of the roots, $\frac{3}{7}$, is rational. This is the heart of the method.
- **Geometry produced arithmetic.** We never searched for cubes; a tangent line found them.

What's really going on

The chord–tangent method

Take a cubic curve, such as $y^3 = x^3 + 9$ or the more standard form $y^2 = x^3 + Ax + B$. If you know two points on it with rational coordinates, the line through them meets the curve in exactly one more point, and that point also has rational coordinates (by the argument in *Looking back*). If you know only one point, use the tangent there instead. Either way, old rational points produce new ones.

Remarkably, this "chord–tangent" construction can be made into a way of *adding* points on the curve, and the rational points then form a *group*, an algebraic structure you will study in your first year. Louis Mordell proved in 1922 that this group is generated by finitely many points: all the rational points, however many there are, can be built from a finite set using chords and tangents. His proof is a version of Fermat's descent from Section 1.5.

Elliptic curves today

Curves like these are called *elliptic curves* (not because they are ellipses, but because of a historical link with the arc length of an ellipse). They are everywhere in modern mathematics.

- **Fermat's Last Theorem.** In the 1980s Gerhard Frey noticed that a solution of $x^n + y^n = z^n$ would produce a very strange elliptic curve. Andrew Wiles proved in 1994 that such strange curves can't exist, and so proved Fermat's Last Theorem.
- **The Birch and Swinnerton-Dyer conjecture.** How many generators does the group of rational points need? A conjecture made in the 1960s, using some of the earliest computer experiments in mathematics at Cambridge, predicts the answer in terms of a function built from counting points modulo primes. It is one of the seven Millennium Prize Problems.
- **Cryptography.** Adding points on elliptic curves over finite number systems is easy, but undoing the addition seems to be extremely hard. That asymmetry protects much of today's online communication.

Sums of like powers

The smallest cube that is a sum of three positive cubes is $6^3 = 3^3 + 4^3 + 5^3$; the tangent method found $20^3 = 7^3 + 14^3 + 17^3$, which is not a multiple of it, and repeating the method keeps producing new identities (although some involve negative numbers). How many kth powers does it take to make another kth power? Euler conjectured in 1769 that you need at least k kth powers to make another kth power. For fourth and fifth powers he was wrong: in 1966 Leon Lander and Thomas Parkin found

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5$$

by computer search, and in 1988 Noam Elkies used elliptic curves to find counterexamples for fourth powers.

The thinking skill

Geometry solving arithmetic

To find whole-number or rational solutions of an equation, think of it as a curve and look for lines through known points.

The algebra behind it is simple: if a polynomial with rational coefficients has all but one of its roots rational, the last root is rational too. Translating between pictures and equations in this way is one of the most powerful habits a mathematician can have.

Try this next

1. **Rational points on a circle.** The line through $(-1, 0)$ with gradient t meets the circle $x^2 + y^2 = 1$ again. Show that the other point is $\left(\dfrac{1 - t^2}{1 + t^2}, \dfrac{2t}{1 + t^2}\right)$, and use this to produce infinitely many Pythagorean triples.

2. **A chord that misses.** The points $(1, 2)$ and $(2, 1)$ lie on $x^3 + y^3 = 9$. Show that the line through them meets the curve in no other point. Where has the "third point" gone? (Look up the *point at infinity*.)

3. **Repeat the trick.** Find the tangent to $y^3 = x^3 + 9$ at $\left(\frac{17}{7}, \frac{20}{7}\right)$ and the third point where it meets the curve. (A computer algebra system helps; notice how quickly the numbers grow.)

4. **Taxicab numbers.** $1729 = 1^3 + 12^3 = 9^3 + 10^3$. Find the third point where the line through $(1, 12)$ and $(9, 10)$ meets the curve $x^3 + y^3 = 1729$, and deduce a new identity of the form $p^3 = q^3 + r^3 + s^3$.

Further reading

- **Joseph Silverman & John Tate, *Rational Points on Elliptic Curves*** (Springer). Written for undergraduates; Chapter 1 needs nothing beyond A-level.
- **Avner Ash & Robert Gross, *Elliptic Tales*** (Princeton University Press). A popular account leading up to the Birch and Swinnerton-Dyer conjecture.
- **Simon Singh, *Fermat's Last Theorem*.** How elliptic curves came to solve a 350-year-old problem.

What We Have Learned

At the start of this chapter we asked three questions about whole numbers, and promised a fourth theme running underneath them: how to prove things. It is time to see how far nine STEP questions have taken us.

1. How many primes are there, and how are they spread out?

In Section 1.1 we proved that there are infinitely many primes without finding a single new one. We built an infinite family of numbers, the Fermat numbers, that never share a factor, so each one needs a prime of its own. It is a proof about *structure*, not size.

In Section 1.2 we went further and showed that the product of all the primes up to n is less than 4^n. That was our first *quantitative* statement: it limits how densely the primes can be packed. Taking logarithms, it says $\sum_{p \le n} \ln p < 1.39\, n$, while the Prime Number Theorem says the true value is about n. Our STEP-level argument was already in the right ballpark.

What we have *not* done is find exactly how the primes are spread out. That is controlled by the Riemann zeta function, and making it precise is the Riemann Hypothesis, the most famous open problem in mathematics. You now know enough to understand what it is about.

2. How can you prove that a number is not a fraction?

We found three quite different ways.

- **Take the smallest one** (Section 1.3). If $\sqrt{2}$ were a fraction, there would be a smallest positive integer k making $k\sqrt{2}$ whole; but then $(\sqrt{2}-1)k$ would be a smaller one. The same argument handled cube roots.
- **Trap an integer between 0 and 1** (Section 1.4). If e or e^2 were a fraction, multiplying by a factorial would produce a whole number lying strictly between 0 and 1. This is the idea behind Hermite's and Lindemann's proofs that e and π are transcendental.
- **Find an impossible divisibility** (Section 1.5, exercise 3). Showing there is no rational point on $x^2 + y^2 = 3$ is the same as showing $\sqrt{3 - x^2}$ is never rational when x is.

And we learned how much remains unknown: nobody can yet prove that $e + \pi$ is irrational.

3. Which equations have whole-number solutions?

Sometimes the answer is *none*, and Section 1.5 showed Fermat's way of proving it: any solution would produce a smaller one, for ever. Working modulo a well-chosen prime told us which divisibility to exploit.

Sometimes the answer is *yes, and here is a machine for producing them*. In Section 1.6 the complex numbers turned sums of two squares into a multiplication rule. In Section 1.9 a tangent line to a cubic curve produced $20^3 = 17^3 + 14^3 + 7^3$.

And sometimes the right move is to *change the number system*. The Gaussian integers made sums of two squares easy. But Section 1.7 gave a warning: in a system that looks perfectly reasonable, numbers can factorise into "primes" in two different ways. That warning, ignored in 1847, led to the invention of ideals and modern algebra.

Section 1.8 added a twist of its own: the fractions, which seem far more numerous than the whole numbers, can be listed one by one. Infinity has more structure than it first appears.

4. A toolkit of proof

Here are the methods you have used in this chapter. Every one of them will appear again and again in a university mathematics course.

Method	Where you used it
Induction and strong induction	Fermat numbers (1.1); products of primes (1.2)
Well-ordering, minimal counterexample	$\sqrt{2}$ and $\sqrt[3]{2}$ (1.3); listing fractions (1.8)
Infinite descent	Impossible equations (1.5); coprime pairs (1.8)
Proof by contradiction	Throughout, especially 1.3 and 1.4
Working modulo n	1.5; 1.7
Bounding and squeezing	1.2; 1.4
Building a bijection	1.8
Constructing a counterexample	Non-unique factorisation (1.7)
Changing the setting	Complex numbers (1.6); geometry (1.9)

If you look back at the questions, you will notice that the STEP examiners often gave you the first step: "by considering the binomial expansion…", "by considering the complex numbers $2 + \mathrm{i}$ and $10 + 11\mathrm{i}$…". At university that help gradually disappears, and *finding* the right first step becomes the skill. The best preparation is exactly what you have been doing: when you finish a problem, ask *why* the hint worked, so that next time you can think of it yourself.

Where next

Several threads from this chapter continue elsewhere in the book.

- The approximation of e by fractions returns in Section 2.9, and the Riemann zeta function in Section 2.4, where $\zeta(2) = \pi^2/6$.

- Roots of unity, which appeared in Lamé's failed proof, are the subject of Chapter 5, and Gauss's construction of the regular 17-gon, promised in Section 1.1, is Section 6.2.
- The four-square version of the identity in Section 1.6 appears in Section 7.8, via quaternions.

If you want to go further before university

- **Harold Davenport, *The Higher Arithmetic***: the best single introduction to elementary number theory.
- **Aigner & Ziegler, *Proofs from THE BOOK***: Chapters 1–4, 8 and 17 connect directly with this chapter.
- **Joseph Silverman, *A Friendly Introduction to Number Theory***: covers almost everything here, and more, at a gentle pace.
- **Marcus du Sautoy, *The Music of the Primes*** and **Simon Singh, *Fermat's Last Theorem***: the human stories behind the mathematics.

Number theory has a habit of drawing people in. Its questions are simple enough to ask on the first day, and deep enough to occupy a lifetime. If some of these sections have left you wanting to know what happens next, that is exactly the point.