Touch Points in Mathematics

Introduction

This short book grows out of the belief that there is extraordinarily interesting mathematics that lies just outside the scope of what is normally covered in high school algebra. What is intriguing about that is that anyone who has taken such a course actually has the keys to understanding some pretty exciting stuff.

The four chapters here begin with "touch points"—topics you may already have seen (although we summarize them here as well). Each chapter deals with a specific remarkable piece of work. The presentation here is a little different than an ordinary math course or even a popularization, because it is mathematics from the point of view of the people who created it. The goal is to give sufficient detail, so that the reader can appreciate what people actually accomplished, but without going too far from the vocabulary of the touch points.

The chapters also show different ways mathematical work can be important. The first chapter is about a specific application area that affects many people's daily lives. The second and third chapters are less applied, but are cases of extraordinary solutions of fundamental problems with many consequences. The fourth chapter presents a monumental piece of work by many people leading to a surprisingly powerful result, with another famous problem as a spin-off.

Four chapters seems like enough concentration to ask of a reader, and—as you'll see—these topics turn out to fit together in interesting ways.

My hope is that the book can be useful from several points of view:

- Most fundamentally I hope this is an interesting story, made more so by presenting the actual mathematics involved.

- Along with the story I hope there is enough detail so that you can understand the kinds of things that mathematicians do. As a high school student it took me a while to get there.

- Perhaps irrationally, I hope this can be useful to some people who have found mathematics hard to get into. It seems to me that one of the problems people have with mathematics is that the vocabulary and notation make it seem like a strange world that only a few people can hope to penetrate. In fact once you get past the definitions, what remains is problem-solving like anything else. The examples here are extraordinary pieces of work, but it is work that can be understood and appreciated for what it is.

Chapter 1: Modular Arithmetic, War, Peace, and Quantum Mechanics

Background

Modular arithmetic is a simple concept with applications everywhere. This chapter first looks at what it does for commerce on the internet. The story then leads straight into quantum computing, so we end up with a window on that as well.

To get things going, we begin with a quick review. Much of this is probably familiar, but it covers what is needed for the rest of the text.

Even if you haven't seen it before, modular arithmetic is a simple idea that turns out to have very many uses. (One indication of usefulness is that every programming language includes modular arithmetic.) It's best to start with an example to introduce the notation. The equation

$$10 = 3 \pmod 7$$

means nothing more than 3 is the remainder when you divide 10 by 7. The number 7 in this case is called the "modulus".

Modular arithmetic is just arithmetic done with remainders. Going back to our example, we know that 0, 1, 2, 3, 4, 5, and 6 are all of the possible remainders when you divide by 7. The remarkable thing with modular arithmetic is that we can do all the same operations with those 7 numbers that you can do with the full set of ordinary numbers.

Addition and multiplication are straightforward as you can see from the following examples

$3 + 6 = 9 = 2 + 7 = 2 \pmod 7$

and $3 \times 6 = 18 = 4 + 14 = 4 \pmod 7$

In other words with addition and multiplication you do the arithmetic as usual, and then find the remainder mod 7. The result is another number between 0 and 6.

Subtraction and division need a little more discussion. Subtraction undoes addition, for example $(3 + 5) – 5 = 3$. Furthermore with ordinary numbers we can rewrite that same equation as $(3 + 5) + (-5) = 3$ (because $(3 + 5) + (-5) = 3 + (5 + (-5)) = 3 + 0 = 3$). So subtraction by 5 is just addition by the number that takes 5 to 0 (namely -5).

Subtraction works the same way for modular arithmetic, but the number that takes 5 to 0 is actually 2--because $5 + 2 = 7 = 0 \pmod 7$. And we see that works, because $(3 + 5) + 2 = 3 + (5 + 2) = 3 + 0 \pmod 7$. So addition by 2 undoes adding 5.

In fact, for any of the numbers n = 1, …, 6 we can see that that what undoes addition by n is just adding the number 7 – n—because $n + (7- n) = 7 = 0 \pmod 7$. Standard terminology for this is to say that 7 – n is the "additive inverse" of n (mod 7). For modular arithmetic, subtraction by n just means adding n's additive inverse 7 – n. (Note that 0 is obviously its own inverse.)

Division follows exactly the same logic, except that we now want to undo multiplication. For ordinary numbers we know that $(3 \times 5) \div 5 = 3$, and we know that division by 5 is the same thing as multiplication

by 1/5, where 1/5 is the number that multiplies 5 to the number 1. (Note that we are only concerned with non-zero numbers now—you can never divide by 0.)

That same idea works for modular arithmetic as well--to undo multiplication by 5, you need to multiply by the number takes 5 back to 1 (mod 7). That number turns out to be 3, since 5 x 3 =15 = 1 mod 7. Following the same terminology we used for addition we say 3 is the "multiplicative inverse" of 5 mod 7.

Overall 3 and 5 (3 x 5 = 15), and 2 and 4 (2 x 4 = 8) are multiplicative inverses of each other mod 7. Also 6 is its own inverse (6 x 6 = 36 = 1 + 35). For modular arithmetic, division by a non-zero number n is just multiplication by the multiplicative inverse of n.

For addition, there was a very simple formula for additive inverses: 7 – n for all positive numbers n. For multiplication, finding the multiplicative inverse is not so easy, but interestingly enough the method to do it goes back to Euclid. Anyone interested in the details should take a look at Appendix 1. For now the main point is that every non-zero element has a multiplicative inverse, and there is an efficient way to find it.

With that, we know how to do all four operations of arithmetic on the numbers 0, 1 ,2, 3, 4, 5, 6 mod 7.

Before leaving this example we need one more bit of terminology—the notion of "groups". A group is just a collection of elements with one operation (think addition or multiplication) and where there is an identity that leaves every element alone (i.e. 0 for addition, 1 for multiplication) and every element has an inverse that takes it back to the identity. We just saw that mod 7 there are two natural groups:

- The additive group with 7 elements (0, 1 ,2, 3, 4, 5, 6) where 0 is the identity and each non-zero element n has the inverse 7 – n.
- The multiplicative group with 6 elements (1, 2, 3, 4, 5, 6) where 1 is the identity and the inverse is calculated as described in Appendix 1.

Going one step further we can say that everything we have just discussed works not just for 7 as the modulus, but for any prime number p as the modulus. In case you don't remember, a prime number is a number only divisible by one and itself. 6 = 2 x 3 is not prime, but 7 = only 1 x 7 is prime. In particular for every prime number p, the (p-1) non-zero elements mod p form a group under multiplication.

The reason we singled out prime numbers p for the modulus is that a non-prime modulus introduces complications for the multiplicative group. The problem is that for the non-prime case, a non-zero element can multiply to 0. For example 3 x 5 = 0 mod 15. Exactly the same thing can happen for any number that divides the modulus, so for non-prime modulus values, the multiplicative group is defined to exclude all numbers that have a prime factor in common with the modulus. If you look at the argument in Appendix 1 you will see that for each of the remaining (non-excluded) numbers, an inverse can be calculated in the same way that it was done for the prime modulus case.

Finally—to end this summary—it turns out to be useful to know the number of elements in the multiplicative group for the simplest non-prime case, where the modulus m is a product p x q of two primes. Consider 15 = 3 x 5. For the multiplicative group we need to exclude all multiples of 3 < 15 (3, 6, 9, 12) then all multiples of 5 < 15 (5, 10) and finally 0. So the total number of elements left in the multiplicative group mod 15 is

15 – 4 [multiples of 3] – 2 [multiples of 5] -1 [for 0] = 8.

The general case works in exactly the same way. The total number of excluded values will be (p-1) multiples of q [the pth multiple is the modulus itself] plus (q-1) multiples of p plus 1 for 0. Therefore the total number of values mod p x q with multiplicative inverses is (just simplifying terms):

$$p \times q - (p-1) - (q-1) - 1 = p \times q - p - q + 1 = (p-1) \times (q-1).$$ [a nice simple formula]

Thus far this is all just theory, but the punch line (if you don't already know it) is more than you might expect.

Virtually all commerce on the internet depends on modular arithmetic in the m = p x q case. The reason is that the security of internet business rests on encryption, and for the most fundamental form of encryption on the net—encoding and decoding are done with powers mod m = p x q!

The security of the system in fact rests on the difficulty of finding the prime factors p and q when you only know the product p x q. In the surprisingly-relevant movie Sneakers, the plot turns on a secret black box that can decode everything. That scenario wasn't completely crazy. Someone who wakes up one morning with an especially good idea for factoring numbers into pairs of primes could still break a fair amount of everything.

Things you can do:

- Find additive and multiplicative inverses mod 11.

- Write a program to calculate the Greatest Common Divisor using the method in Appendix 1. If you'd like a little more, you can extend the program to do multiplicative inverses. For that you can either unwind the GCD algorithm as in the Appendix, or you can look up the Extended Euclidean Algorithm—which adds bookkeeping to compute the inverse as you go.

Encryption

With all that as introduction we can now get into the meat of the story. We will talk first about encryption, then how we got to where we are, and finally some surprising bits about the future.

The first thing to say about encryption is that in general terms it is something familiar to just about anyone. You take a message and transform it some way so that the receiver can undo the transformations and read it.

With traditional encryption, as has existed in various forms for centuries, encoding and decoding are two sides of the same thing—knowing the encoding rule tells you how to undo it and decode. The encoding rule can be some explicit formula for letter substitution, or it can be settings on a piece of equipment that codes and decodes messages, or it can be key values for a software program that does the same thing. However in all cases it is critical that both ends know the rule.

Figure 1: Engima

The picture above shows the famous Nazi Enigma machine from World War II—which serves as an example of what can go right and wrong with traditional encryption. On the plus side, the rotors, rings, plugboard and other settings of the basic version of the Enigma allowed for about 10^{23} possible settings for the rules. This was on the face of it far more than enough to be make decryption impossible by searching through all the settings.

On the minus side, however, was the problem of communicating key information to widely distributed sites, including even submarines. (Each side needed the same settings for encryption to match with decryption.) That was managed using codebooks of daily settings and elaborate handshake protocols at the start of every session. Ultimately it was failures of this operational side of the system, including stolen code books and human failings of Nazi operators, that reduced the scale of the problem to the point where brilliant people like Alan Turing could break the system. (See here for a concise retelling of that wild story—much more dramatic than the movies made about it.)

Given that the internet is far more distributed and unprotected than the Nazi lines of command, it is natural to wonder what changed and when to make billions of dollars of internet commerce possible.

Public Key Systems

The answer very specifically was the 1976 invention of public key encryption. With public key encryption, the encryption and decryption processes are NOT the same. In fact the encryption process can be made public without compromising the decryption process that remains secret. So you can publish an encryption key that anyone can use to send you a secure message—hence the name public key.

That sounds nice but mysterious, so we'll get to the specifics right away. The system we describe is the RSA system (named for its inventors Rivest, Shamir, and Adelman) from 1978.

The RSA system is the one mentioned earlier based on the difficulty of factoring large numbers that are products of two primes. So we start with a modulus m = p x q. However the first step is a little surprising. Instead of working with the modulus m, we use the fact that we know p and q individually and work with the multiplicative group mod (p-1)(q-1). We choose a number e that is any member of the multiplicative group mod (p-1)(q-1). As described before, that just means that e is not divisible by any prime dividing (p-1) or (q-1).

Now whatever the message may be, it can be expressed as a series of numbers, for example the numbers corresponding to the way the message would be stored as data. After all, everything in a computer is expressed as numbers. (To be concrete, the letter "C" in ASCII is stored as binary 01000011, which is the ordinary decimal number 67.) So we can think of any message as broken up into a series of numbers N. The formula for the encoding is now surprisingly simple. The encoding for each message N is just N^e mod m. Note that only m = p x q and e are exposed, not p or q individually.

The question now is how to decode. The key fact that makes this go is that for every element n in the multiplicative group mod m, $n^{(p-1)(q-1)}$ = 1 mod m. We prove that using what we already know about the multiplicative group mod m in Appendix 2. This is actually a special case of a theorem about elements in any group—any element raised to the "total number of elements in the group" power is the identity. For this case we saw earlier that (p-1)(q-1) is the number of elements in the multiplicative group (mod m). This form of the result is referred-to as the "little Fermat" theorem.

What the relation $n^{(p-1)(q-1)}$ = 1 mod m gives us, is that powers n^s depend only on s mod (p-1)(q-1). To see that suppose we add some multiple t of (p-1)(q-1) to s. Say t = r(p-1)(q-1).

$n^{s+t} = n^s \times n^t = n^s \times n^{r(p-1)(q-1)} = n^s \times (n^{(p-1)(q-1)})^r = n^s \times 1^r = n^s$ (all equations mod m)

Therefore for any number s

$$n^s = n^{s \bmod (p-1)(q-1)} \text{ (mod m).}$$

Using this we can now decode. N^e (mod m) was the encoded value of the message N we started with. The exponent e was a member of the multiplicative group mod (p-1)(q -1). From Appendix 1, we know how to find inverses for any modular multiplicative group (that is assuming we know the modulus, which in this case means knowing p and q). Let's call d the calculated inverse of e mod (p-1)(q-1).

Because d is the inverse of e, e x d = 1 mod (p-1)(q -1), so finally we have

$$(N^e)^d = N^{e \times d} = N^{e \times d \bmod (p-1)(q-1)} = N^1 = N \text{ (again all mod m)}$$

In other words, we just need to find d as the multiplicative inverse of e (once and for all), raise the encoded message to the d power (mod m), and we're done. Again notice that the only reason we can do this is that we know not just the product m = p x q, but p and q individually. Otherwise we couldn't find (p-1)(q-1) and the inverse d. Publishing m and e isn't enough to do the decoding, you need the still secret values of p and q (with d as the calculated result).

An example makes this clearer. For this we take m = 115 = 5 x 23, i.e. p = 5 and q = 23.

For encryption and decryption we need to look at the multiplicative group mod (p-1)(q-1) = 88. As encryption exponent e we take the number 7, which has no factors in common with 88. Next, using the

algorithm of Appendix 1, we find that the multiplicative inverse of 7 mod 88 is 63. We confirm by noting 7 x 63 = 441 = 1 mod 88. So the decryption exponent d = 63.

We now publish the modulus m = 115 and the encryption exponent 7 as the encoding scheme for messages to us. The decryption exponent 63 stays secret. Note once again that the security rests on knowing the prime factors 5 and 23 of m. That is what lets us compute d via the method of Appendix 1.

We can try that it works by encrypting the letter "C" using the ASCII value 67.

Encoding is 67^7(mod 115) = 33. For decoding we calculate 33^{63}(mod 115) which is in fact 67.

Now that we know the basics, there is quite a bit to say about using the system. First start with security. Factoring numbers into primes is an old problem with a long history of algorithms. In general people classify difficulty of algorithms by how the number of steps grows with the size of the problem (in this case the size of the modulus m to be factored). As a rule, easier problems grow as some power of the size. Harder problems cannot be bounded by any specific power. Since they grow so rapidly in computing requirements, it is easy to choose a size beyond any conceivable computing equipment. Even today, with 40 years of RSA motivation, the factoring time for the best algorithms still grows faster than m^n for any fixed power n.

Notice this isn't a proof, just the state of the art. However, many of such harder problems have been shown to be equivalent to each other, so the grounds for faith are wider than just a single problem.

Second, it should be noted that public key algorithms are computationally more intensive than traditional encryption. Raising to powers takes more computing resources than addition and multiplication. So public key systems are often used for key exchange prior to traditional encryption (using software analogues of Enigma). Interestingly enough one of the first proposed public key systems (the so-called knapsack) only used addition and multiplication, but was broken by Shamir (the S of RSA) in 1982.

Third, the RSA system and many other public key systems also offer the advantage of so-called "signatures". For signatures the order of encoding and decoding steps is reversed, so that the owner of a public key takes a chosen text N and encodes it as N^d using his secret decoding exponent d. If the receiver then raises the result to the published value e, he will see the decoded text that could only have been sent by someone knowing the private value d. This constitutes a signature identifying the sender.

Fourth, as a practical consideration, in a real system the individual RSA messages contain appended pseudo-random bits, so that values of encoded messages will not be repeated in case of duplicated text. The same sort of procedure deals with values of N that share factors with the modulus m. Operationally this just means that certain bits of each decrypted N value are not included in the decoded message.

Just as a matter of interest, in the late 70's public key encryption seemed like such a miracle that people couldn't quite believe it. At Bell Labs for instance they wanted to use it to encrypt so-called "signaling information" (who is calling whom) sent between switches and transmitted over the air via microwave. People were sufficiently uneasy about the whole idea that they talked about setting up their own (highly-paid) black hat organization to try to break RSA!

In 2002 the RSA people received the ACM Turing Prize (a kind of Nobel prize for computer science). Ron Rivest's Turing Prize talk is available online and gives an interesting view of how they thought about what they were doing. Also, in the last few years, public key encryption has been front page news for both good reasons and bad. On the bad side, most so-called ransomware uses RSA encryption to hide a person's own computer files until the ransom is paid. Ron Rivest's comment was that he felt "sort of like a mother whose son was brainwashed and left to become a jihadist in Syria". On the good side, bitcoin and blockchains—the current revolution in finance—are fundamentally based on public key encryption.

Things you can do

Put together a public key encryption system with a couple of primes of your choice. Try a few example encoding/decodings. If you're up for it, you can actually program a working system.

Quantum computing and Shor's algorithm

We've now had 40 years of RSA without significant challenges to its security. However recently notice was given by the NSA that it "must act now" to prepare for the days when RSA and other equivalent algorithms would no longer be secure. So again the question is what has happened to change the picture?

The answer, as indicated by the title, is quantum mechanics, or more specifically quantum computing. Strangely enough, public key encryption has emerged as the poster child for quantum computing. The capabilities of quantum computing are in fact so well-matched to RSA and other such systems, that public key encryption is the best example yet of a case where a quantum computer is vastly superior to a conventional computer.

Peter Shor's quantum factoring algorithm of 1994 is theoretically capable of breaking RSA. The question is when someone will be able to build a big enough quantum computer to do it. To date the biggest number factored with it is 21 = 3 x 7. (Other much larger numbers—e.g 56153 = 233 x 241--have been quantum-factored by other methods, but those are special cases not relevant to encryption.)

The goal for the remainder of this note is to say enough about quantum computing and the factoring algorithm, so that you can appreciate what it is about quantum computing that works in this case. This is also as good an example as any for a look at quantum computing. As will be clear from the example, the reason that quantum computing can solve the intractable problem at the root of RSA is not that a quantum computer is simply faster. It is that in Shor's algorithm the operations needed for factoring are carefully matched to what quantum hardware can deliver.

That being said, quantum computing is weirder but less complicated than you might think. Just as the basic unit of ordinary computing is a bit, the basic unit of a quantum computer is a qubit. Ordinary bits are easy to explain—each bit is independent and is either 0 or 1, so the state of a collection of bits is its binary value. Qubits are not independent—they can be "quantum entangled", and the system as whole is not in a single state, but in a combination ("superposition") of states with differing probabilities. Clearly an example is necessary.

8 bits in an ordinary computer can represent can represent one of 2^8 = 256 different numbers (depending on which bits are 0 or 1), and that single number says everything there is to know about the state of those 8 bits.

In a system with 8 qubits, each one of those 256 numbers is a possible value for the 8 qubits, and the system overall can be in any combination of those 256 values with associated probabilities. Think about it. Instead of having eight bits that can be 0 or 1, we have 256 system values with a probability of occurrence attached to each. So instead of a system where each state is defined by 8 binary values, we have one where each state is defined by a 256-tuple of probabilities.

To put numbers on this, we can get a lower bound by looking at how many values we can get with 256 binary bits. (That amounts to requiring all of the probabilities in a given state to be equal.) By the same logic we started with, that number is $2^{(2^8)}$ or about 1.58 x 10^{77}. So a system with relatively few qubits can represent a huge application. A single quantum transformation can do a lot of work, because it takes one combination of qubits values and probabilities to another such combination, subject to quantum rules about which transitions are allowed.

However there is a critical restriction—you can't know the state of qubits directly. You have to "measure" them. And when you measure (or evaluate) a qubit or set of qubits, you will get a single one of the possible values for that qubit or set, and the states inconsistent with the measured values are lost. In other words, when you measure a lot of information gets lost. We'll see in a minute what that means.

Even though many different types of hardware are currently under development for quantum computing, all of them rely on the same set of properties of quantum entanglement. So—remarkably enough—there is a mathematical theory of quantum computing that is independent of whether the qubits are atoms or photons or something else. There is even a standard language for representing logical operations on qubits. In general quantum states are represented by vectors, and the allowed quantum operations correspond to so-called "unitary" matrices. A good reference to get started is "Programming Quantum Computers" by Johnston, Harrigan, and Gimeno-Segovia.

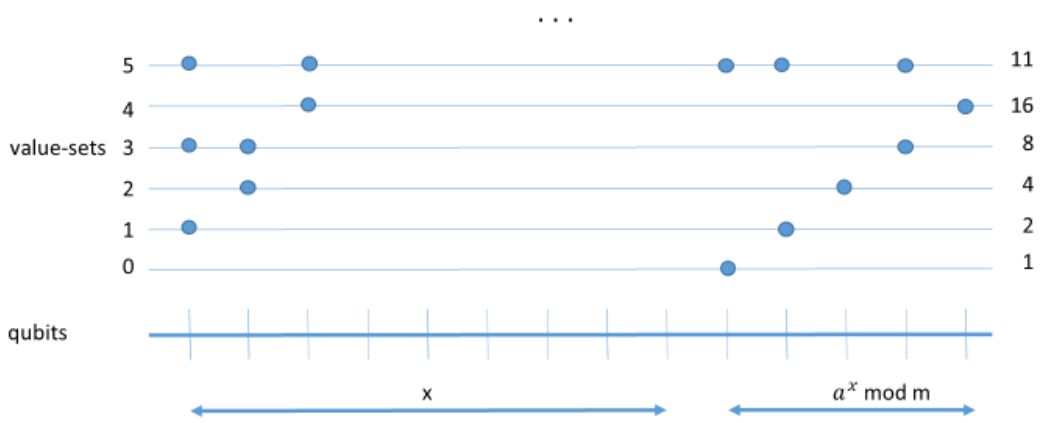
With that we can start on the Shor's algorithm. We'll assume we're given a number m that is known to be a product m = pq of two primes. Our job is to find p and q. (Recall that once we know p and q, we know (p – 1)(q – 1) and with that we can find the inverse mod (p – 1)(q – 1) for the encoding exponent e.)

The first thing to note is that the quantum part of the algorithm doesn't find the prime factors directly. What it does is choose a random element "a" mod m, and then it calculates the smallest number k such that a^k = 1 mod m. (We can assume a is relatively prime to m, because we can check for common divisors.) From that k value Shor uses a conventional (non-quantum) method to find p and q. We'll go through the quantum algorithm first, then the conventional part.

In order to find the smallest number k such that a^k = 1 mod m, the algorithm uses a superposition of qubit value-sets where each value-set represents both a single value x and the corresponding value of a^x (mod m). Figure 2 shows what this looks like in terms of qubits for factoring m = 21. The hatched lines at the bottom represent individual qubits—in this case there a 9 qubits on the left representing values of x, and 5 more on the right representing values of a^x (mod 21). For the horizontal lines

representing value-sets, the x values are shown on the left, a^x (mod 21) on the right. The dots show which qubits are active for the value-set (this is just ordinary binary arithmetic). To be clear, Figure 2 only shows the first few value-sets. As there are 9 qubits to represent x, there are a total of 2^9 = 512 equally likely value-sets.

Figure 2: State for Schor's Algorithm
(m = 21, a = 2)



[In case you are wondering, we have 9 qubits for x because Shor's algorithm requires state values for x = 0, 1 ,2, …, (2^N – 1) where N is determined by the inequality $m^2 < 2^N < 2 \times m^2$. For m = 21, this means $441 < 2^N < 882$. So 2^N = 512 and N = 9. For a^x (mod 21) the argument is simpler—the maximum value is 20 and $2^4 < 20 < 2^5$, so we need 5 qubits to do it.]

The algorithm starts with all qubits null. The system is first initialized by standard quantum operations with the 512 input values for the x variable, each with equal probability. A second quantum operation then extends all of those initial values in parallel to include the corresponding a^x (mod 21). That puts the system in the state shown by Figure 2.

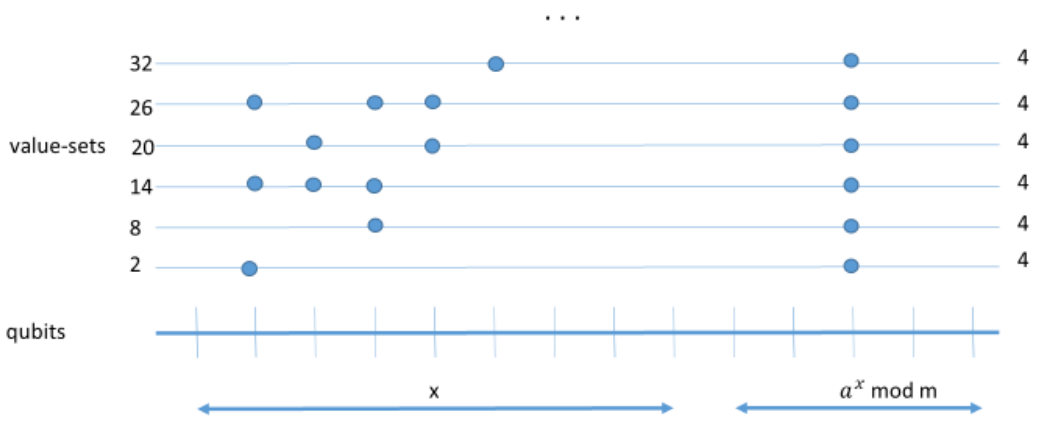
On the face of it, once we reach the configuration of Figure 2 it seems we ought to be done. We have values of x with corresponding values of a^x, so it seems we ought to be able to tell when a^x first gets to 1. The problem is that we're talking about qubits, so we can't see the state information directly. Even though we can see the answer in the figure, there is nothing we can evaluate that will tell us anything about it. The hard part of the algorithm is getting the information out!

The first thing we do is evaluate the second set of qubits (corresponding to the values a^x mod m). When we do that, one of the possible values gets chosen at random. That value of itself is no use, but once that value is determined, all the value-sets incompatible with that value go away and we're left with the x values that correspond to the single measured value of a^x mod m. Since the sequence of values of a^x mod m repeats over and over again as x increases, the differences between successive remaining x values will be precisely the smallest number k such that a^k = 1 mod m. (We keep

multiplying by a until we get back the same value we started with (mod m). That means the result of our multiplications was 1 mod m.)

Figure 3 shows what happens to the states in Figure 2 when we evaluate the last 5 qubits and get 4. We are now left with only those value-sets with a^x = 4 (mod 21), which reduces the number of superposed, equally-likely value-sets from 512 to 86.

Figure 3: State after third quantum operation
(m = 21, a = 2, a^x mod m evaluates to 4)



We still can't see the value-sets, so we have to come up with a way to get the difference information out. For this we perform another standard qubit operation—the so-called Quantum Fourier Transform—which adds up the x and a^x values with appropriate coefficients so that characteristics of repeated patterns drop out. (See Appendix 3 for more detail.) With that operation, the x-variable probabilities get rewritten so that the highest probability values give estimates of k.

We now evaluate the x-variable qubits. One estimate isn't enough, but multiple quantum calculations eventually get the result. In estimating the performance of this algorithm, Shor was able to put bounds on the number of trials necessary to get the value k, so for the quantum computation stage we're done.

To end this we need to turn the equation a^k = 1 mod m into possible factors for m. That's done via a conventional mathematical argument.

a^k = 1 mod m is the same as a^k -1 = 0 mod m, and if k is even we can rewrite that as

($a^{k/2}$ - 1) ($a^{k/2}$ + 1) = 0 mod m (using the fact from algebra that for any number x, x^2–1 = (x – 1)(x +1)).

Now since m is a product of two primes p and q, this means that unless m itself divides one factor or the other, we will have one prime dividing the first and the other prime dividing the second. (The product of the two primes divides the product of the two factors, so the primes will be either together in one of the two factors or apart.)

If the primes are apart, the argument in Appendix 1 contains an efficient algorithm to find the greatest common divisor of two numbers, and we can use that algorithm applied to m and the two factors individually to pull the primes out of each factor. So for k even and m not dividing either factor we are done.

To finish completely one shows that when we selected the variable "a" there was a greater than 50-50 chance that the exponent k would be even and the same is true for the probability that m doesn't divide one of the two factors. This means that the algorithm only needs to be applied a few times to have a high probability of success. And with that we're completely done.

To review, the algorithm works like this:

1. Choose a random value a mod m, and then setup the initial qubit states as in Figure 2. For now we are looking for the minimum k such that a^k = 1 mod m.

2. Evaluate the second set of qubits (for a^x mod m). The remaining x-variable values will each differ from the previous by the k value we're looking for.

3. Apply the Quantum Fourier Transform.

4. Repeat steps 1, 2, and 3 as necessary to determine the value of k.

5. Repeat all 4 previous steps with a new value for "a" if k is odd or m divides ($a^{k/2}$ - 1) or ($a^{k/2}$ + 1).

With this algorithm the overall realtime is estimated to grow as $N^2(\log N)^3$. That estimate makes the improvement over traditional methods very clear--all classical algorithms grow faster than any fixed power! [Recall that N is the number of values in the Shor algorithm and satisfies $m^2 < 2^N < 2 \times m^2$. That means N grows as the log of m, so realtime grows even slower than $m^2(\log m)^3$.]

The only barrier to doing this is the number of qubits required. Even for a modulus as small as 21 we saw we needed 9 + 5 = 14 qubits for a full application of the algorithm—and 14 is quite a large number with current technology. A realistic decryption scenario would involve a 2048-bit modulus and 4000 qubits—so no one is going to break RSA tomorrow. But there sure are a lot of people trying!

There are in fact many active areas of work:

- Because Shor's algorithm is specific to RSA, there are other public-key encoding schemes under development that would not be broken by known quantum algorithms.

- Naturally enough there is also work on possible quantum-based encryption.

- Already today there are applications of quantum computing to enhance the security of existing encryption. For example, by distributing "entangled" qubits, it is possible to check that transmitted bits have not been tampered with.

So in the end this is actually an important case of work in progress. RSA was a great idea, but it won't last forever, and no one knows what the successor will be.

Things you can do

Try Peter Shor's approach to factoring for m = 15 = 3 x 5. First choose a number a and find the first integer k with a^k = 1 mod m. (You'll have to do the quantum part manually by calculating powers of a mod m.) Then check if Shor's approach works for those values of a and k. Otherwise try another value of a until one works.

m = 15 is quickly done. The next case, 35 = 5 x 7, is a little more work.

You can even program Shor's whole algorithm. That's not so difficult to do. Finding k is straightforward to program even though it gets realtime-intensive. With a program you can try factoring larger numbers.

Appendix 1: Greatest Common Divisors and Multiplicative Inverses

This section gives an efficient algorithm to find the inverse for any member of the multiplicative group mod m, prime or not. The argument also shows how to find greatest common divisor of any two numbers.

We start with the greatest common divisor. As a convenient notation, let GCD(A, B) be the greatest common divisor of A and B, i.e. the largest number dividing both. Let's assume A > B and divide A by B to get A = BQ + a remainder R < B. The first and most important result is

GCD(A,B) = GCD(B,R)

Since R = A – BQ, any divisor of A and B clearly divides R and of course B. So the remaining question is whether there is anything extra that divides B and R but not A. But A = BQ + R says that anything that divides B and R divides A.

We can now repeat the same procedure with B and R, dividing the larger by the small and taking the remainder. At each stage the remainder is strictly smaller than the larger of the two numbers we started with, so the process ends after at most be B steps.

There are two ways the process can end (otherwise we can just continue).

1. We reach a stage where the remainder is 1, in which case that is the greatest common divisor, and the two numbers have no factors in common.

2. We reach a stage where the remainder is 0 for two numbers M and N. In that case, we have

   GCD(A,B) = GCD(N,M) = the smaller of M and N.

We now have a simple, efficient procedure to find greatest common divisor. What does this say about multiplicative inverses mod m?

Suppose we pick any number a < m that has no prime factors in common with m. Then if we go through the procedure to find the greatest common divisor of a and m, we will reach the stage where we are working with two numbers N and M where N = MQ +1 (as a and m have greatest common divisor = 1).

The point now is to recognize that N and M are either equal to a or m, or else remainders of a previous divisions. In the remainder case, say C = QD + M, then M = C -QD, and if we replace M in the equation by C-QD, we can move back to the values in the previous iteration. If we unwind all the way back, we get an equation that only involves the values a and m from the beginning. In fact once we collect terms we will end up with an equation that looks like r x a + s x m = 1 for some numbers r and s. If we reduce that mod m, we get r x a = 1 mod m, so we have found the inverse.

An example makes this clearer. Let's consider 8 mod 21.

Step 1: 21 = 2 x 8 + 5

Step 2: 8 = 1 x 5 + 3

Step 3: 5 = 1 x 3 + 2

Step 4: 3 = 1 x 2 +1

This gives the GCD of 8 and 21 = 1 (i.e. they are relatively prime.) To find the inverse of 8 mod 21 we work backwards.

The 2 in Step 4 is the remainder from Step 3, so we have

3 = (5 – 3) + 1.

The 3's are now 8-5 from Step 2, so we get

(8 – 5) = (5 – (8 – 5)) + 1.

Finally the 5's are now 21 – 2 x 8, so we end up with an expression with multiples of 8's and 21's:

(8 – (21 – 2 x 8)) = ((21 – 2 x 8) – (8 – (21 – 2 x 8)) + 1.

Simplifying we get

8 – 21 + 2 x 8 = 21 – 2 x 8 - 8 + (21 - 2 x 8) + 1 or

3 x 8 – 21 = 2 x 21 – 5 x 8 + 1 or

8 x 8 = 3 x 21 + 1.

This says 8 x 8 = 1 mod 21 (or 8 is its own multiplicative inverse) which checks since 8^2 = 64 = 3 x 21 + 1.

This ends the discussion of the algorithm. However we should be clear that the intent thus far has been to show that the GCD calculation also gives the inverse. For practical calculation of inverses there is a reorganized and somewhat more complicated version of the algorithm that avoids the two stages we just saw, i.e. working down to the GCD and then back up to the inverse. You can see this "Extended Euclidean Algorithm" here.

Appendix 2: $n^{(p-1)(q-1)}$ = 1 mod m, for any number n prime to m, where m = p x q is a product of 2 primes.

The first thing to prove is that n^k = 1 (mod m) for some number k ≤ (p-1)(q-1). This is just a counting argument. Consider the set S = {n, $n^2, n^3, \ldots, n^{(p-1)(q-1)}$} of the first (p-1)(q-1) powers of n. Since there are (p-1)(q-1) elements in the set S and also (p-1)(q-1) elements in the multiplicative group, either 1 is in the set S, or some other value occurs twice, say as n^i=n^j mod m. If j is the bigger power, then multiplying both sides by the inverse of n^i mod m we get 1 = n^{j-i}. So in both cases we have n^k = 1 mod m for some number k ≤ (p-1)(q-1). If k = (p-1)(q-1) we're done.

Otherwise assume k is the smallest value with n^k = 1. Now define a new set S = {1, n, … , n^{k-1}}, and consider the sets aS = {a x 1, a x n, … ,a x n^{k-1})} where a is any element in the multiplicative group mod m. Since S includes 1, every element in the multiplicative group belongs to one of these sets aS. Furthermore the sets aS are either identical or distinct, since an intersection of aS and bS means there are elements a and b with a x n^i = b x n^j, and that implies a = b x n^{j-i}, so aS = bS. If l = the number of distinct sets aS, then (since each aS has k elements) counting elements gives k x l = (p-1)(q-1). That now gives us what we want since

$$n^{(p-1)(q-1)} = n^{kl} = (n^k)^l = 1^l = 1 \text{ mod m.}$$

Notice this argument can be applied to any group to show that any element raised to the "total number of elements in the group" power is the identity.

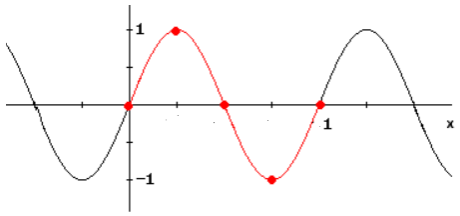
Appendix 3: The Fourier Transform in Shor's Algorithm

To understand the Fourier transform in Shor's Algorithm, we need to start with a simpler concept—a Fourier series. In this first part we will be dealing not with quantum computing, but ordinary real numbers.

Suppose we have some periodic function f with period 1. A Fourier series for f expresses it as a sum of sine and cosine functions, as we will show in a minute.

The most basic sine and cosine functions with period 1 are $\sin(2\pi x)$ and $\cos(2\pi x)$, as indicated in Figure A1.

Figure A1: $y = \sin(2\pi x)$



In fact for every positive whole number n, the functions $\sin(2\pi nx)$ and $\cos(2\pi nx)$ are all periodic with period 1, because they repeat the behavior in Figure A1 n times as x goes from each whole number to the next.

The crucial point is that it turns out that every "reasonable" periodic function with period 1 can be expressed uniquely as a series using these functions $\sin(2\pi nx)$ and $\cos(2\pi nx)$.

Another way to say the same thing is that we can write our function f as a sum of sines and cosines:

$$f(x) = \sum_{n=0}^{\infty} a_n \sin(2\pi nx) + b_n \cos(2\pi nx)$$

This is called a Fourier series for the function f. The ∑ symbol here just means "sum". The a_n's and b_n's are constants specific to the function f, and they reflect the importance of each particular sine or cosine term in describing f. There are formulas from Calculus that show how to calculate the values of the a_n's and b_n's. The bigger the value of particular a_n's and b_n's, the more important those particular $\sin(2\pi nx)$ and $\cos(2\pi nx)$ terms are in understanding the function as a whole. Think about notes on a piano—there's a fundamental frequency that says which note it is and a lot of other overtones that give character to the sound.

You can think of a "Fourier transform" as the process that takes a given function and calculates all the a_n's and b_n's of the Fourier series. It "transforms" the function f into the Fourier series that represents it. The reason behind this whole approach is that for many applications it is easier to understand the behavior of a function from its Fourier series than by plotting points.

To see the reasoning behind Shor's algorithm we need to look at a slightly more general formula. Instead of assuming our function has period 1, let's say we have a function g with period p. In that case we get a slightly more complicated-looking formula—we divide the sine and cosine arguments by p:

$$g(x) = \sum_{n=0}^{\infty} a_n \sin\left(\frac{2\pi nx}{p}\right) + b_n \cos\left(\frac{2\pi nx}{p}\right)$$

If we compare the two series for f and g, there's something interesting. Since f has period 1, it also has period p by definition. So we should be able to write f in the format we've given for g. In fact we can match up the two expressions for f if we write it this way:

$$f(x) = \sum_{\substack{n=0, \\ p\ divides\ n}}^{\infty} a_n \sin\left(\frac{2\pi nx}{p}\right) + b_n \cos\left(\frac{2\pi nx}{p}\right)$$

Every term in that second expression for f is exactly the same as a term in our first formula. The point is that even though we started with a formula for functions of period p, we can tell if a particular function actually has the shorter period 1 by looking at the a_n's and b_n's. Specifically, all the a_n's and b_n's are 0 for n's not divisible by p, if and only if the function really has <u>period divided by p</u>.

The idea behind the using the Quantum Fourier Transform is that if we play our cards right, we can calculate a_n's and b_n's for the function implied by Figure 3, and then use them to find the period we're looking for. In general terms it works like this…

The function we're going to transform here is the function f(x) = a^x (mod m) with the x-values that remain after the second quantum operation—that is in situation of Figure 3. All of those f(x) values are the same, 4 in Figure 3. The corresponding x values differ from each other by k; in Figure 3 that difference is 6. For all the other (missing) x values, we can take f(x) to be 0. That defines f(x) every integer from 0 to 2^N-1 (0 to 511 in Figure 3). We can define f for all integers just by repeating these values over and over again. Another way to say the same thing is that for any integer x, f(x) = f(x mod 2^N).

From the way we defined it, f is periodic with period 2^N. However we also know that f has the smaller period k—the value we're looking for—because the function repeats for every k values of x. Following the example, we do our Fourier transform with period p = 2^N(512 in Figure 3). That gives us the a_n's and b_n's for the function f in the form:

$$f(x) = \sum_{n=0}^{\infty} a_n \sin\left(\frac{2\pi nx}{2^N}\right) + b_n \cos\left(\frac{2\pi nx}{2^N}\right)$$

If we were in the case we started with, we would find k easily because the only non-zero entries would have n = a multiple of $\frac{2^N}{k}$. (To be clear, k is the new period; $\frac{2^N}{k}$ is what the original period of $2^N$was <u>divided by</u> to end up with a period of k.) Here we don't quite have that case, because the calculations are different and k may not be a power of 2, so that it doesn't exactly divide 2^N. However Shor showed—by some effort—that the situation is good enough. (The strange inequality for the number of x-values $m^2 < 2^N < 2 \times m^2$ is part of it.) When we look at the a_n's and b_n's we'll see clear peaks at n =

multiples of $\frac{2^N}{k}$, and smaller values for everything else. Since k = $2^N \div \frac{2^N}{k}$, that's how the Quantum Fourier Transform determines k.

We still have to get that number out! In the Quantum Fourier Transform the a_n's and b_n's are combined into single values p_n representing a combined magnitude of a_n and b_n. Then all the p_n's are divided by a constant, so that they sum to 1. Finally the probabilities of the x-values are changed so that each number n = 0, 1 ,2, …, $(2^N - 1)$ occurs with probability p_n. Since the p_n's come from a_n's and b_n's, multiples of $\frac{2^N}{k}$ occur with the highest probabilities.

So we evaluate the x-variable. We won't always get the right answer, but if repeat the quantum calculation over and over again we'll know k. We just have to do enough test calculations to see the pattern. Shor calculates how many tests we'll need for his estimate of the overall efficiency of the algorithm, and with that we're done.

As a graphic indication of why quantum computing matters for this application, the reference mentioned earlier gives the following chart comparing performance of the Quantum Fourier Transfer (QFT) with the corresponding classical computation (FFT) as the size of the problem increases. Because of the difficulties we've seen in getting data in and out, the QFT can't be used for all FFT applications, but where it can be used—as in this case—the difference is dramatic.

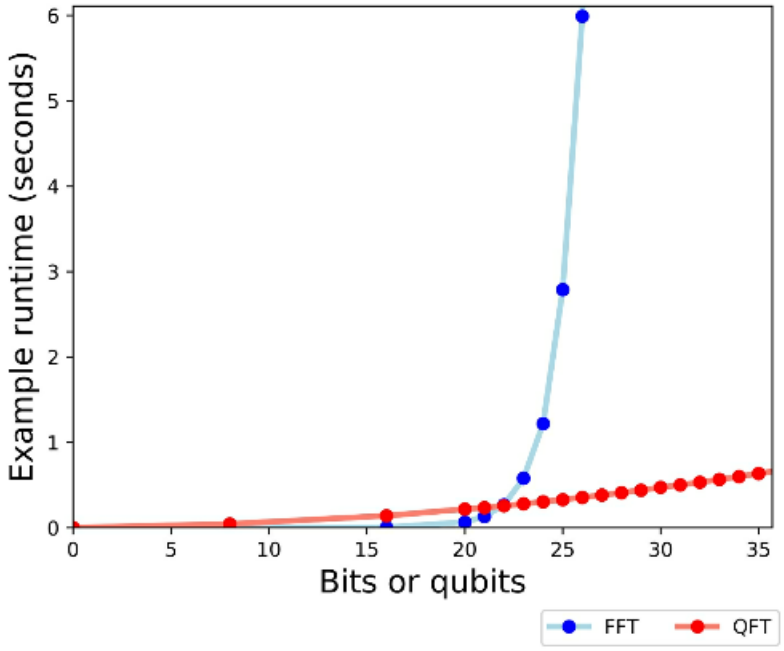


Figure 7-20. Time to compute QFT and FFT on a linear scale

Chapter 2. Riemann and Primes

Background

This chapter follows-on from the last one—and also introduces one of the most famous problems in mathematics. We've just seen that the security of the RSA encryption system rests upon the difficulty of factoring large numbers into primes. However, thus far we've talked only generally about factoring, and that's the connection here.

Factoring algorithms ultimately come down to searching for primes given some conditions on what to look for. As a result, the performance of many such algorithms depends on what we can say about the distribution of primes—the better our information about primes, the better the performance bounds we can give for the algorithms. The distribution of primes is also important for many other discrete applications for the same kind of reason.

The crucial event for what we know about the distribution of primes was a short (eight-page!) paper published by Bernard Riemann in 1859. That gave an asymptotic formula for the number of primes less than any given value x, as well as an indication of how much variation to expect around that value. Not everything in that paper was proved by Riemann himself, but his line of argument kept generations of mathematicians busy. And the story ends with a fundamental bit still unresolved.

Also, interestingly enough, we'll see there is much in Riemann's paper that grows out of topics you've seen before.

Review of Series

We begin by recalling a little of what we know about series. A mathematical series is a finite or infinite sum of terms following some pattern.

One useful example is the so-called geometric series. That is defined by an initial value "a" and a ratio "r" so that nth value for the sum is

$$S_n = a + ar + ar^2 + \ldots + ar^{n-1}.$$

It's easy to sum this series since

$$rS_n - S_n = ar^n - a \text{ (all the other terms cancel).}$$

$$\text{Then } S_n(r\text{-}1) = a(r^n - 1), \text{ so that}$$

$$S_n = \frac{a(r^n - 1)}{(r-1)}$$

As an example we can look at $S_7 = 2 + 2 \times 5 + 2 \times 5^2 + \ldots + 2 \times 5^7$. Using the formula we can immediately write down the result as

$$S_7 = \frac{2(5^8 - 1)}{(5-1)} = \frac{2(390625 - 1)}{4} = 195312$$

Geometric series turn up everywhere, very frequently in finance, and we'll see several here.

If r is less than 1, then it makes sense to talk about the value of the infinite series

$$S = a + ar + ar^2 + \ldots \text{ or equivalently } S = \sum_{n=0}^{\infty} ar^n$$

For each value of n, the sum $S_n = \frac{a(r^n - 1)}{(r-1)}$, and since r < 1 the nth power of r goes to 0 as n increases.

So the whole infinite sum S has the value $\frac{a(0-1)}{(r-1)} = \frac{a}{(1-r)}$.

(Note that the ∑ symbol, if it looks strange, is actually simple. It just means "sum" and it tells you where to start and end for the values that you're adding.)

This is the first example most people see of an infinite series with a specific, computable value.

For $S = \sum_{n=0}^{\infty} 2\,x\,(\frac{1}{5})^n$, for example, the sum is $\frac{2}{(1-\frac{1}{5})} = \frac{2}{\frac{4}{5}} = \frac{5}{2}$.

We'll return to geometric series a little later, but for now we'll talk about another simple series that is frequently discussed—the so-called harmonic series. That series looks like this

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots \text{ or equivalently } \sum_{n=1}^{\infty} \frac{1}{n}$$

The interesting thing about this series is that it is provably bad—it doesn't have a well-defined value, and that's easy to show. We just have to group the terms:

$$1$$

$$+\frac{1}{2}$$

$$+\frac{1}{3} + \frac{1}{4} > \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$+\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} > \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

etc.

This pattern shows that the sum will be larger than any given number if you take enough terms. Specifically,

To get a sum value of 1 takes 2^0 or 1 term

To get a sum value > 2 takes 2^2 or 4 terms

To get a sum value > 3 takes 2^4 or 16 terms

To get a sum value > 4 takes 2^6 or 64 terms

Each successive sum is 4 times the last, so that in general—to get a sum value > n takes $2^{2(n-1)}$ terms.

That means the sum is unbounded and cannot converge.

However if we replace $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$ with squared values $1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots$ then something quite different happens. There is again a simple proof.

$\sum_{n=1}^{N} \frac{1}{n^2} < 1 + \sum_{n=2}^{N} \frac{1}{n(n-1)}$ ($n(n-1)$ is always less than n^2, so the reciprocal is greater)

But $\frac{1}{n(n-1)} = \frac{1}{n-1} - \frac{1}{n}$, so the right hand side is

$$1 + \sum_{n=2}^{N} \left(\frac{1}{n-1} - \frac{1}{n}\right) = 1 + 1 - \frac{1}{N}$$ (since all but first and last terms in the sum cancel).

Therefore $\sum_{n=1}^{N} \frac{1}{n^2} < 2$, and the series converges.

In fact the actual value, proved by Euler in1734, turns out to be $\frac{\pi^2}{6}$!

Riemann's paper

We're now ready to look at Riemann's results. His fundamental insight was that the key to understanding the distribution of prime numbers is the so-called Riemann zeta function:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

We already know a little about this function, because we've just been looking at ζ (1)—which diverged and ζ (2)—which was better-behaved and equal to $\frac{\pi^2}{6}$. A natural question is where it stops converging. In fact the harmonic series is just barely over the edge, because the series for $\zeta(s)$ turns out to converge for <u>every</u> real number s > 1. The proof in the Appendix is simple, but requires Calculus.

We'll now begin looking at the relationship between the ζ function and primes.

First of all, we can factor our expression for $\zeta(s)$ into separate factors per prime. That's because (as we know) every integer can be factored uniquely as a product of prime powers, e.g. 140 = 2^2 * 5 * 7. So if we write things out in the following form:

$$\zeta(s) = \left(\sum_{n=0}^{\infty} \frac{1}{{p_1}^{ns}}\right)\left(\sum_{n=0}^{\infty} \frac{1}{{p_2}^{ns}}\right)\left(\sum_{n=0}^{\infty} \frac{1}{{p_3}^{ns}}\right) \quad \ldots$$

we'll get all the $\frac{1}{n^s}$ values from combinations of terms from each of the sums.

Each of those sums, however, is now a geometric series with a =1 and r = $\frac{1}{p_i{}^s}$. so we can sum the series using our geometric series formula to get

$$\zeta(s) = \left(\frac{1}{1-\frac{1}{p_1{}^s}} \right)\left(\frac{1}{1-\frac{1}{p_2{}^s}} \right)\left(\frac{1}{1-\frac{1}{p_3{}^s}} \right) \quad \ldots$$

which can be rewritten as

$$\zeta(s) = \left(\frac{1}{1-p_1{}^{-s}} \right)\left(\frac{1}{1-p_2{}^{-s}} \right)\left(\frac{1}{1-p_3{}^{-s}} \right) \ldots$$

Finally, we take the logarithm* of both sides (recalling that log (ab) = log a + log b and log (1/a) = - log a)

$$\log \zeta(s) = - \log (1 - p_1{}^{-s}) - \log (1 - p_2{}^{-s}) - \log (1 - p_3{}^{-s}) \ \ldots$$

We're far from done here, but with this expression it's obvious that the behavior of the ζ function is intimately tied up with properties of primes.

To turn that vague connection into a formula, there are three steps:

1. We need to understand more about the behavior of the ζ function.
2. We need an explicit statement of the relationship of the ζ function to the distribution of primes.
3. We need to put it all together to get a formula for the distribution of primes.

Step 1. More about the ζ function

To start, we need to emphasize that for Riemann's analysis we need to look at $\zeta(s)$ as a function of a complex number s. So instead of looking at ζ as a function on the real line, we regard it as a function on the complex plane of numbers a +ib. (There is a quick summary of complex numbers in Appendix 1 of the next chapter.)

The first thing to check is convergence. We saw earlier that the series for $\zeta(s)$ converges for every real number s > 1. What about for complex numbers s = a + ib?

The result is actually almost the same. For complex numbers we need to look at

$$\sum_{n=1}^{\infty} \frac{1}{n^{a+ib}} = \sum_{n=1}^{\infty} \frac{1}{n^a}\frac{1}{n^{ib}}$$

And for that we get

$$\sum_{n=1}^{\infty} \frac{1}{n^{a+ib}} \leq \sum_{n=1}^{\infty} \frac{1}{|n^{a+ib}|} = \sum_{n=1}^{\infty} \frac{1}{|n^a|}\frac{1}{|n^{ib}|} = \sum_{n=1}^{\infty} \frac{1}{n^a}$$

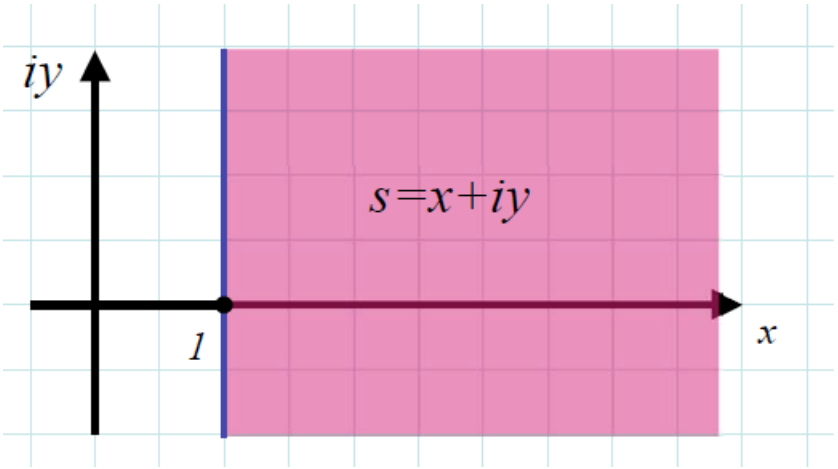
* All logarithms here are natural logarithms to the base e.

The last step follows from Euler's equation e^{ix} = cos x + i sin x, which means $|e^{ix}| = 1$. Here, however, we need to replace e^{ib} by n^{ib}. For that case we have

$$n^{ib} = e^{(\log n)^{ib}} = e^{i\,(\log n)b} = \cos(\log n)b + i \sin(\log n)b,$$

so $|n^{ib}| = 1$ in this case also. (We'll see other uses for Euler's equation later.)

Since the series $\sum_{n=1}^{\infty} \frac{1}{n^a}$ converges for every real-valued s > 1, this says that the series for $\zeta(s)$ over the complex numbers converges for every complex number with real part > 1. Otherwise stated, the function $\zeta(s)$ converges for every complex number s in the half-plane defined by Re(s) > 1, as shown by the colored area below.



Now we get to a first surprising part. Riemann combines the zeta function with the classical gamma function (explained in a minute) to extend the zeta function—which blew up at ζ (1)—to the entire complex plane!

We say more about the gamma function in the Appendix, but all you really need to know is that the gamma function fills in between integers for the factorial function n!. Until now we've thought of the factorial function as just for positive integers n. The gamma function is defined for all complex numbers and matches the factorial function on positive integers. For historical reasons, the definition is such that Γ (n) = (n-1)! (instead of the more obvious n!) For our purposes it's just a standard way to extend n! to the complex numbers.

Riemann multiplies the zeta function ζ(s) by the gamma function Γ, (s-1), and a power of π to get a new function with remarkable properties. To keep track of it, we'll call that new function g. The specific formula is $g(s) = \pi^{-½s}\, \Gamma(½s)(s-1)\, \zeta(s)$.

Riemann shows that g is well-defined in the whole complex plane. (The factor (s-1) cancels the bad behavior at 1 we saw with the harmonic series.)

Going back to the formula for g, we can solve for ζ(s) and get

$$\zeta(s) = \frac{\pi^{½s} g(s)}{(s-1)\Gamma(½s)}.$$

Since everything on the right side is defined everywhere the other functions are, this extends the definition of ζ(s) to the whole complex plane—except for a singularity at s =1, which is explicitly like 1/(s-1).

Furthermore this function g is turns out to be symmetrical in an extraordinary way:

$$g(s) = g(1-s)$$

That means the behavior ζ(s) for Re(s) < 0 is determined by the values we started with for Re(s) > 1. Specifically if we write out the function g on both sides we get

$$\pi^{-½s}\ \Gamma(½s)(s-1)\ \zeta(s) = \pi^{-½(1-s)}\ \Gamma(½(1-s))\ s\ \zeta(1-s) \text{ or}$$

$$\zeta(s) = \left(\frac{\pi^{½(s)}\ \Gamma(½(1-s))\ s}{\pi^{½(1-s)}(s-1)\Gamma(½s)\ (s-1)}\right)\ \zeta(1-s)$$

If Re(s) < 0 then Re(1-s) will be > 1, so the formula tells you how the old values determine the new ones.

We can summarize what we know about ζ(s) by dividing up the complex plane as follows:

- For Re(s) > 1, values are determined from the series for ζ function, which is well-defined and always of positive magnitude.

- For Re(s) < 0 , values are determined from the first case, using the formula for exchanging s and 1-s we just saw. It turns out that ζ(s) is 0 for s = -2, -4, -6 etc., because ζ(s) cancels out singularities of the gamma function at those points. There are no other zeroes of ζ(s) in this region.

- The remaining vertical strip with 0 <= Re(s) <= 1 is called the "critical strip". We don't know anything about the behavior of ζ(s) there yet, but as the name indicates that will be critical for the distribution of primes. For now all we can say is that the zeroes of ζ(s) are symmetric around the line Re(s) = 1/2 (because of g(s) = g(1-s)) and also symmetric about the real line (because complex conjugation works term-by-term, ζ(conjugate(s)) = conjugate(ζ(s)). So ζ(s) = 0 implies ζ(conjugate(s)) = 0 also.)

So we've taken a zeta function that looked like it was going to useful only for Re(s) > 1, and found a way to extend it to the rest of the complex plane. Moreover we found a functional equation relating values on the right side of the complex plane with values on the left. And then finally we decided that the part of the zeta function's domain we care most about is the vertical strip between Re(s) = 0 and Re(s) = 1, where originally it seemed that the function wasn't defined at all!

Step 2. Relating the ζ function to the distribution of primes

The starting point here is the formula we found earlier for log $\zeta(s)$ in terms of primes:

$$\log \zeta(s) = -\log\left(1 - {p_1}^{-s}\right) - \log\left(1 - {p_2}^{-s}\right) - \log\left(1 - {p_3}^{-s}\right) \ \ldots$$

Riemann was able to take it a step farther, using a formula from Calculus for log (1 – x). Specifically (for |x| < 1) the formula says

$$\log(1-x) = -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \ldots$$

We can apply that directly to each of the terms $\log(1 - p_i^{-s})$ and we get

$$\log \zeta(s) = -(-p_1^{-s} - \frac{1}{2}p_1^{-2s} - \frac{1}{3}p_1^{-3s} - \ldots) - (-p_2^{-s} - \frac{1}{2}p_2^{-2s} - \frac{1}{3}p_2^{-3s} - \ldots) - \ldots$$

All the minus signs cancel out, and we end up with the formula:

$$\log \zeta(s) = (p_1^{-s} + \frac{1}{2}p_1^{-2s} + \frac{1}{3}p_1^{-3s} - \ldots) + (p_2^{-s} + \frac{1}{2}p_2^{-2s} + \frac{1}{3}p_2^{-3s} - \ldots) + \ldots$$

But that is actually simpler than it looks. It's just:

$$\log \zeta(s) = \sum_{all\ positive\ integers\ n} \sum_{all\ primes\ p} \frac{1}{n} p^{-ns}$$

This is a sum over all powers of primes, and for each term there is a coefficient of 1/(the power) that multiples p^{-ns}.

These coefficients that match the powers of primes turn out to work miracles.

In order to get at the distribution of primes, Riemann defines a new function J(x) that basically adds up those coefficients for all numbers less than x. Specifically J(x) starts out = 0 at zero. After that it jumps at prime powers—by 1 for a prime, by ½ for a squared prime, by 1/3 for a cubed prime, etc. The jumps occur at prime powers and are the same as the values of the coefficient function just mentioned.

J(x) turns out to have three remarkable properties. (Since this is the a crucial part of Riemann's paper, we give considerable detail in the Appendix to this chapter, but it requires Calculus.)

- First, the double sum

$$\log \zeta(s) = \sum_{all\ positive\ integers\ n} \sum_{all\ primes\ p} \frac{1}{n} p^{-ns}$$

turns out to be determined by the function J in a simple and intuitive integral equation!

- Second that equation for log ζ as a function of J can be inverted to give J in terms of log ζ. That gives the core relationship between the ζ function and the counting of primes.

We give the expression here for completeness (although it's complicated and involves Calculus):

$$J(z) = \frac{1}{2\pi i} \int_{a-\infty i}^{a+\infty i} \frac{\log \zeta(s)}{s} z^s ds$$

However the details are not important. What is important is that we have an explicit formula to compute the prime-counting function J in terms of zeta-function $\zeta(s)$.

- Third, J is obviously counting primes in an odd sort of way, but in fact it's better than that. If we define the function $\pi(x)$ to be the number of primes less than x (this is standard notation), then by definition—think about it—we have

$$J(x) = \pi(x) + \frac{1}{2}\pi\left(x^{1/2}\right) + \frac{1}{3}\pi\left(x^{1/3}\right) + \frac{1}{4}\pi\left(x^{1/4}\right) + \ldots$$

And that one can be inverted too!

$$\pi(x) = J(x) - \frac{1}{2}J\left(x^{1/2}\right) - \frac{1}{3}J\left(x^{1/3}\right) - \frac{1}{5}J\left(x^{1/5}\right) + \ldots$$

The nth coefficient here is $\frac{\mu(n)}{n}$, where μ(n) is the so-called Möbius function—which is 0 if n is divisible by a prime squared, otherwise 1 if the number of prime factors is even, -1 if the number of prime factors is odd. The overall technique used to reverse the roles of J and π is known as "Möbius inversion".

This means that we can approximate $\pi(x)$ as closely as we want using terms in J(x).

With that we now have a way to compute $\pi(x)$—the number of primes less than x—in terms of the log of the ζ function. It remains to turn that into a formula.

Step 3. A formula for the distribution of primes

The formula comes from looking in more detail at our function g(s) = $\pi^{-½s}$ Γ(½s)(s-1) ζ(s) . The point is that we can get an entirely new way of thinking about g by looking at its roots.

Suppose for the sake of argument that g was a polynomial. In that case we would know exactly how to express g as a function of its roots.

When all the roots of a polynomial are real, we know from algebra classes that the polynomial splits up as a product of individual factors corresponding to the roots. If we call the roots r_i, a polynomial p(x) looks like:

$$p(x) = (\text{constant } C)\,(x - r_1)\,(x - r_2)\ldots(x - r_N).$$

It is an important fact (called the Fundamental Theorem of Algebra and proved by Gauss in 1799) that over the complex numbers every polynomial (with real or complex coefficients) can be broken down like this. Otherwise stated, all the roots of every polynomial can be expressed as complex numbers. So over the complex numbers every polynomial p(x) looks like

$$p(x) = C\,(x - r_1)\,(x - r_2)\ldots(x - r_N)$$

for possibly complex roots r_i. To go one small step farther, if none of the roots are zero, we can divide every term by $(-r_i)$ and modify the constant C to compensate. That gives the following form for p(x):

$$p(x) = (\text{new constant } D)\left(1 - \frac{x}{r_1}\right)\left(1 - \frac{x}{r_2}\right)\ldots\left(1 - \frac{x}{r_N}\right)$$

Plugging in x = 0, shows that the constant D is just p(0).

So if g were a polynomial of degree N, we could write it as $g(s) = g(0)\left(1 - \frac{s}{r_1}\right)\left(1 - \frac{s}{r_2}\right)\ldots\left(1 - \frac{s}{r_N}\right)$.

In actual fact g isn't a polynomial, since Riemann produces a formula for number of zeroes of the zeta function with imaginary part < T, and that formula implies that the number of zeros is infinite. But we still get the next best thing!

It turns out that the infinite product of the $(1 - \frac{s}{r_n})$ terms actually converges for all values of s, that g(0) ≠ 0, and finally that g can be expressed as the infinite analog of our polynomial formula, namely

$$g(s) = g(0)\,(1 - \tfrac{s}{r_1})\,(1 - \tfrac{s}{r_2})\,(1 - \tfrac{s}{r_3})\,\ldots$$

As for the zeroes r_i, we know that g(s) is well-defined on the whole complex plane, and it is non-zero for Re(s) > 1 (since the ζ function is) and also for Re(s) < 0 (by the functional equation). So all of its zeroes are in the critical stip. Furthermore because $\pi^{-½s}$ and Γ(½s) are non-zero on the critical strip—and (s-1) just cancels the blowup of ζ (1)—we can say that the zeroes of g(s) are precisely the zeroes of the zeta function in the critical strip.

We now have two very different expressions for g(s):

$$g(s) = \pi^{-½s}\,\Gamma(½s)\,(s-1)\,\zeta(s) \text{ from the definition, and}$$

$$g(s) = g(0)\,(1 - \tfrac{s}{r_1})\,(1 - \tfrac{s}{r_2})\,(1 - \tfrac{s}{r_3})\,\ldots \text{ from the zeroes of g}$$

Equating the two we get

$$\pi^{-½s}\,\Gamma(½s)\,(s-1)\,\zeta(s) = g(0)\,(1 - \tfrac{s}{r_1})\,(1 - \tfrac{s}{r_2})\,(1 - \tfrac{s}{r_3})\,\ldots$$

or

$$\zeta(s) = \frac{\pi^{½s}}{\Gamma(½s)(s-1)}\quad g(0)\,(1 - \tfrac{s}{r_1})\,(1 - \tfrac{s}{r_2})\,(1 - \tfrac{s}{r_3})\,\ldots$$

Taking logarithms again, we get

$$\log \zeta(s) = \tfrac{s}{2}\log \pi - \log \Gamma(½s) - \log(s-1) + \log g(0) + \left(\sum\nolimits_{roots\ r_i\ of\ \zeta} \log\,(1 - \tfrac{s}{r_i})\right)$$

This all looks rather messy, but the important thing is that we've expressed log ζ(s) as a sum of known functions plus the sum over the roots. And we can use that expression to get a formula for the distribution of primes. Remember that in step 2 we had an equation for the prime counting function J in terms of log ζ(s), namely

$$J(z) = \frac{1}{2\pi i}\int_{a-\infty i}^{a+\infty i} \frac{\log \zeta(s)}{s}\, z^s\, ds$$

We can plug our expression for log ζ(s) into that equation for J, and compute the integral term-by-term. When you do that the result is an <u>explicit formula for prime-counting function J</u>. (We'll first get a form that includes operations from Calculus, but then we'll give a simplified form that does not.)

The basic result is this (we'll talk about the Li function in a minute):

$$J(x) = \mathrm{Li}(x) - \sum\nolimits_{roots\ r_i\ of\ \zeta} \mathrm{Li}(x^{r_i}) - \log 2 + \int_0^\infty \frac{1}{t(t^2-1)\log t}\, dt$$

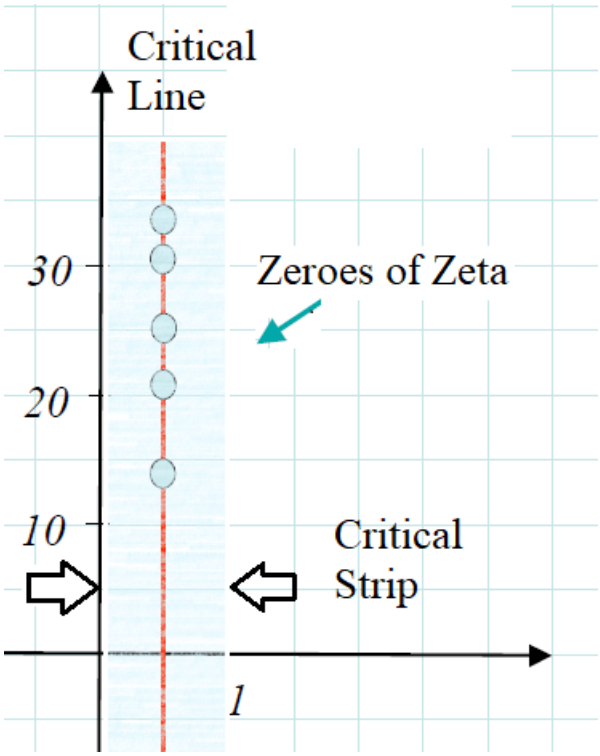
That may look intimidating, but even before explaining the terms this is an amazing result. It is an exact formula for a function that is counting up primes! Somehow all of these terms put together (including the infinite sum) give an exact count of the primes and prime powers used to compute J!

Also in the end it is simple. The last two terms are constants bounded by 1, so they are for practical purposes negligible. Which means we're left with a formula term Li(x) followed by an error term expressed as a sum over zeroes of the zeta function.

The formula term uses the Li function (for "log integral")—that is defined using Calculus by the equation $\mathrm{Li}(x) = \int_2^x \frac{dt}{\log t}$. For those unfamiliar with the integral, however, there is a simpler alternative. Li(x) is asymptotically equal to $\frac{x}{\log x}$ (meaning the ratio of the two functions goes to 1 as x goes to infinity). So Li(x) and $\frac{x}{\log x}$ are basically equivalent formulas for J(x) and hence for the number of primes < x . (We can also get more precise—but more complicated—estimates by using additional terms from Riemann's expression for π(x) in terms of J.)

However even though we have formulas, there is still a big step remaining. For these formula terms to have value we have to show that the other non-trivial term—the sum over the roots of the ζ function in the critical strip—is small relative to the formula term for large x.

Riemann asserted (but didn't claim to prove) that all of the roots should lie on the line Re(s) = ½. This is the so-called "Riemann Hypothesis", and the line Re(s) = ½ is referred-to as the critical line.



The Riemann Hypothesis minimizes the magnitude of the sum over roots, and therefore maximizes the accuracy of π(x) = Li(x) for the number of primes < x. If the Riemann hypothesis is true, the absolute

value of the difference $\pi(x)$ - Li(x) grows as $x^{1/2} \log x$—much slower than Li(x) or $\frac{x}{\log x}$. (This follows since log x grows slower than any positive power of x.)

If the Riemann hypothesis is false and all one can say is Re(s) <= ½ + e, then the error grows as $x^{\frac{1}{2}+e} \log x$. If e < ½ this is still small relative to the formula, but since the change is in an exponent, the error grows significantly.

The Prime Number theorem (that the number of primes < x grows like Li(x) or $\frac{x}{\log x}$) was proved by Hadamard and De La Vallee Poussin in 1896, 37 years after Riemann's paper. To do that, they showed that there are no zeroes of ζ(s) on the line Re(s) = 1 (which of course implies no zeroes on Re(s) = 0 as well). That was good enough for the theorem, but the bound on the sum of roots was still large.

The Riemann Hypothesis has remained as a kind of ultimate challenge to this day.

- The first major result was in 1914 when the British mathematician G. H. Hardy proved there are infinitely many zeroes of the zeta function on the critical line Re(s) = ½.

- In 1942 the Norwegian mathematician Alte Selberg showed that a (very small) positive proportion of the zeroes are on the critical line.

- In 1974 Norman Levinson of MIT (dying of cancer) proved that 1/3 of zeroes are on the critical line.

- In 1989 J. B. Conrey (late of the University of Bristol) improved that to 40%.

- In 2020 the team of Pratt, Robles, Zaharescu and Zeindler made made it 5/12.

And that's where the theoretical work stands now.

There have also been impressive results of computer searches for roots of the zeta function.

Andrew M. Odlyzko:
The 10^{20}-th zero of the Riemann zeta function and 175 million of its neighbors all satisfy the Riemann hypothesis.

Sebastian Wedeniwski:
The first 10^{11} nontrivial zeros of the Riemann zeta function lie on the line Re(s) = 1/2.

There is even a kind of postscript. In addition to the roles of primes for encryption and other discrete applications, there have also been recent applications of the zeta function itself in physics. The distribution of zeroes of the zeta function is apparently related to energy levels in atomic nuclei!

Appendix

The main writeup for this chapter avoids Calculus wherever possible. However much of Riemann's mathematical machinery is based on Calculus, so a next level of detail is provided here. There are three topics referenced from the main text.

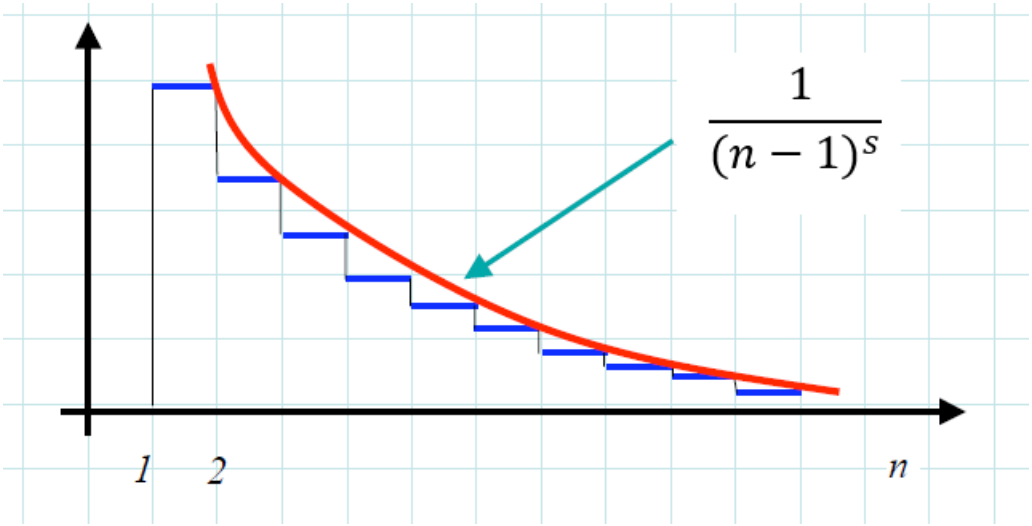
1. $\zeta(s)$ converges for every real number s > 1.

The trick here is that when s is a real number we can look at $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$ geometrically. As shown by the blue lines in the following figure, you can view $\zeta(s)$ as a sum of rectangles, each one unit wide and with a height of $\frac{1}{n^s}$ for each integer n. Superimposed in that picture is the function $\frac{1}{(n-1)^s}$ shown in red. For n >= 2, the red line matches the corners of the blue graph, and everywhere else it sits above it. So if we compare the areas under the blue lines and red lines we get

$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$ < 1 (area of the first rectangle) + $\int_2^{\infty} \frac{1}{(n-1)^s} dn$ (area under the red curve).

For s > 1, that integral is well-defined; it is (by the ordinary power rule) $\frac{1}{(1-s)(n-1)^{s-1}}$, evaluated from 2 to ∞. That is 0 - $\frac{1}{(1-s)(2-1)^{s-1}}$ or simply $\frac{1}{(s-1)}$.

Hence $\zeta(s) < 1 + \frac{1}{(s-1)}$ and the series converges.



2. The role of the gamma function Γ(s) in the functional equation for $\zeta(s)$

The gamma function is defined by the integral $\Gamma(s) = \int_0^{\infty} t^{s-1} e^{-t} dt$. That's less exotic than it looks—it's just a trick with integration by parts. You can easily show $\Gamma(n+1)$ = n $\Gamma(n)$ using integration by parts, and it's not hard to verify $\Gamma(2) = 1! = 1$. Strictly speaking the integral defines Γ for Re(s) > 0,

but Γ gets extended to the whole complex plane (with singularities at negative integers), much as we saw with the zeta function.

The full proof of the functional equation g(s) = g(1-s) out of scope here, but we will show how closely the gamma function is related to $\zeta(s)$.

The first step is to make a change of variables t -> nt in the equation defining Γ. Gamma itself is unchanged, but we get an interesting new equation:

$$\Gamma(s) = \int_0^\infty (nt)^{s-1} e^{-nt} n dt = n^s \int_0^\infty t^{s-1} e^{-nt} dt$$

$$\text{or } \int_0^\infty t^{s-1} e^{-nt} dt = \frac{\Gamma(s)}{n^s}$$

If we now sum both sides over n (and pull the sum inside on the left), we get

$$\int_0^\infty t^{s-1} (\sum_n e^{-nt}) dt = \Gamma(s) \ (\sum_n \frac{1}{n^s}) = \Gamma(s)\, \zeta(s)$$

The sum in the left side is another geometric series, this time with a = e^{-t} and r = e^{-t}. So we have

$$\sum_n e^{-nt} = \frac{e^{-t}}{(1-e^{-t})}, \text{ or multiplying top and bottom by } e^t,$$

$$\sum_n e^{-nt} = \frac{1}{(e^t-1)}.$$

Substituting back in the integral, we end up with a simple expression for the product $\Gamma(s)\ \zeta(s)$:

$$\Gamma(s)\ \zeta(s) = \int_0^\infty \frac{t^{s-1}}{(e^t-1)} dt\ .$$

So the relation of the gamma and zeta functions is surprisingly fundamental. For now this is true for Re(s) > 1, where the zeta function is defined. However there is more to say, since Riemann shows that the integral on the right-hand side turns out to be well-defined for all complex numbers s.

At the time of Riemann's paper, the Γ function was much-studied and known to be well-defined and non-zero except for singularities at non-positive integers. So dividing by $\Gamma(s)$ is defined and gives a value of $\zeta(s)$ everywhere except at those singularity points of Γ (and at s=1 where we know ζ is undefined). In other words with this simple argument we've extended the zeta function from its original domain Re(s) > 1 to almost the entire complex plane! (That is everywhere except the non-positive integers and s = 1.)

It's considerably more complicated to show that the function g(s) = $\pi^{-½s}$ Γ(½s)(s-1) ζ(s) is defined everywhere and satisfies g(s) = g(1-s), but with this example we can already see how to extend the domain of the zeta function well beyond what first seemed possible. And for that the relation between the zeta and gamma functions is key.

(For those with a little more mathematical background, this notion of extending the domain of a function goes by the name of analytic continuation. An "analytic" function is a function that can be represented by a power series in every region where it is defined. That's true for the original zeta function, since the infinite series can be differentiated term-by-term to get the power series. The same is true for the integral definition of zeta that we just introduced here. In any overlap of regions, the

power series must be the same, because all of the derivatives must be. So the extended zeta function we've defined here really is a single analytic function in the whole broader domain.)

3. Fourier transforms and the prime-counting function J(x)

The first step here is to interpret the sum in the equation

$$\log \zeta(s) = \sum_{all\ positive\ integers\ n} \ \sum_{all\ primes\ p} \ \frac{1}{n} p^{-ns}$$

using the prime-counting function J. Recall that J is defined as follows: J(x) starts out = 0 at zero. After that it jumps at prime powers—by 1 for a prime, by ½ for a squared prime, by 1/3 for a cubed prime, etc.

There is one more fact we need before we can put things together. For any prime p and power n,

$$p^{-ns} = s \int_{p^n}^{\infty} t^{-s-1} dt \ .$$

There's nothing fancy about that formula—it's just integration of a power of the variable t—but the implications are immense. It's saying that each p^{-ns} term in our sum for log $\zeta(s)$ is an integral with the same integrand but a different starting point.

With that you can think about the summation process for log $\zeta(s)$ in a different way. Imagine you're walking down the number line until you hit a prime power p^n. When that happens, you start an integral with integrand t^{-s-1} and coefficient $\frac{s}{n}$. Each such integral goes to infinity.

As you continue walking, at any position t there will be a certain number of integrals in progress. The integrands are all the same (t^{-s-1}) and the sum of the coefficients of those integrals is by definition equal to s J(t). That means that adding terms in the double sum is the same as walking to infinity with the current value being integrated always = $sJ(t)t^{-s-1}$. In other words we have the remarkable formula:

$$\log \zeta(s) = s \int_0^{\infty} J(t) t^{-s-1} dt$$

That's already good progress—we're turned a doubly infinite sum into a single integral with a perfectly comprehensible function J. However we're only half-way there. What we really want is the other way around—J in terms log $\zeta(s)$.

For that we need a Fourier transform. We introduced Fourier transforms in Appendix 3 of the previous chapter, where we talked initially about Fourier series. The idea is that you can think of a periodic function f(t) in two ways—either in its original form as a function of time or as a sum of standard wave forms sin nt and cos nt with coefficients a_n and b_n chosen for f so that f(t) = $\sum_n \ a_n$ sin nt + b_n cos nt. The original function f(t) and the set of coefficients $\{a_n, b_n\}$ are equivalent representations of the same function, and the process of moving from one representation to the other is a Fourier transform.

The original function f(t) is said to be in the "time domain" and the representation by coefficients $\{a_n, b_n\}$ is said to be in the "frequency domain". One important generalization of the picture just presented

is that (as we'll see in a minute) the frequency domain components can turn out to be a continuous set, rather than indexed by integers. In that case f(t) is reconstructed as an integral rather than a sum. (This process of working back and forth between frequency and time domains is so important in applications that the 1965 breakthrough algorithm for it, the so-called Fast Fourier Transform, was characterized in 1994 as "the most important numerical algorithm of our lifetime".)

Fourier's theorem is specific about which functions can be transforms of each other. For a function A(x) to be expressed as a Fourier transform of a function B in the form

$$A(x) = \int_{-\infty}^{\infty} B(y)e^{iyx}dy ,$$

the function B must itself be expressible as

$$B(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} A(x)e^{-ixy}dx.$$

(Note that by Euler's equation e^{ix} = cos x + i sin x, so this is actually a continuous version of the Fourier series just mentioned.) The idea here is that if A can be built up out of B, then we can reverse that process to build B from A.

On the face of it none of this seems to have anything to do with J or log ζ, so Riemann had work to do to make it go. The process looks complicated, but it's actually just a change-of-variables trick.

First of all, Riemann chooses to write our basic equation in the following way:

$$\frac{\log \zeta(s)}{s} = \int_0^{\infty} J(t)t^{-s-1}dt .$$

Then he uses the following substitutions in that equation:

i) s = a + iy with a > 0 a real constant and y a real variable

ii) We define a new variable x by t = e^x

iii) A(x) is defined to be the function: $2\pi J(e^x)e^{-iax}$

With those carefully-chosen definitions, everything falls into place. Our equation becomes

$$\frac{\log \zeta(a+iy)}{a+iy} = \int_{-\infty}^{\infty} J(e^x)e^{-(a+iy)x}dx \quad \text{(note that dx = dt/t)}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} A(x)e^{-iyx}dx.$$

So we can use the Fourier transform formula with A(x) as just defined, and B(y) = $\frac{\log \zeta(a+iy)}{a+iy}$.

From the formula we have A(x) = $\int_{-\infty}^{\infty} B(y)e^{iyx}dy$. Plugging in the values for A and B we get

$$2\pi J(e^x)e^{-iax} = \int_{-\infty}^{\infty} \frac{\log \zeta(a+iy)}{a+iy} e^{iyx}dy , \text{ or}$$

$$J(e^x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\log \zeta(a+iy)}{a+iy} (e^x)^{a+iy}dy$$

And with one more change of variable z = e^x we get a rather simple final form

$$J(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\log \zeta(a+iy)}{a+iy} z^{a+iy} dy \text{ , or}$$

$$J(z) = \frac{1}{2\pi i}\int_{a-\infty i}^{a+\infty i} \frac{\log \zeta(s)}{s} z^{s} ds$$

As noted in the text, this is the key result, since it gives an explicit formula for the prime-counting function J in terms of the ζ function. To get a formula for J we plug in our expression for log ζ , namely:

$$\log \zeta(s) = \frac{s}{2}\log \pi - \log \Gamma(\tfrac{1}{2}s) - \log(s-1) + \log g(0) + \left(\sum_{roots\ r_i\ of\ \zeta} \log\left(1 - \frac{s}{r_i}\right)\right)$$

Then we evaluate the resulting integral, term-by-term, to yield (as discussed in the text):

$$J(x) = \mathrm{Li}(x) - \sum_{roots\ r_i\ of\ \zeta} \mathrm{Li}(x^{r_i}) - \log 2 + \int_0^{\infty} \frac{1}{t(t^2-1)\log t}\, dt$$

For more detail on any of this material the primary reference is H. M. Edwards, Riemann's Zeta Function, Dover, 1974. Two interesting and more elementary treatments are http://wwwf.imperial.ac.uk/~hjjens/Riemann_talk.pdf and https://medium.com/@JorgenVeisdal/the-riemann-hypothesis-explained-fa01c1f75d3f .

Chapter 3. The Quadratic Formula and Its Descendants

Background

A beginning algebra course is a combination of different kinds of mathematics. The first part of the course just works through consequences of a single good idea—that a variable can be treated just like a number. Based on that idea we get a method to solve simple (linear) problems as soon as we can write them down. Insight alone can be worth a lot.

However when we hit quadratic equations, there is something different. The simple approach doesn't work anymore, and we need to come up with something extra to make progress. Remarkably there is a complete solution, and we end up with a formula you can use even if you don't remember where it comes from, i.e. that the equation $ax^2 + bx + c = 0$ (for any numbers a, b, and c, $a \neq 0$) has solutions

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

This chapter grows out of the quadratic formula, but it is really about a single extraordinary insight.

The quadratic formula has a long history. Rules for solving quadratic equations go back to Euclid, and were derived independently by many mathematicians in many cultures since then. The current form of the solution with an explicit indication of two solutions came later, in the sixteenth century.

Also in the sixteenth century people were able to go beyond the quadratic formula to find results for higher-order equations, with increasingly complex results. For cubic equations the basic result due to Cardano was the following:

The equation $ax^3 + bx^2 + cx + d = 0$ has the solution

$$x = \sqrt[3]{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right) + \sqrt{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right)^2 + \left(\frac{c}{3a} - \frac{b^2}{9a^2}\right)^3}}$$

$$+ \sqrt[3]{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right) - \sqrt{\left(\frac{-b^3}{27a^3} + \frac{bc}{6a^2} - \frac{d}{2a}\right)^2 + \left(\frac{c}{3a} - \frac{b^2}{9a^2}\right)^3}} - \frac{b}{3a}.$$

This gives a single solution. You can see the general formula that gives all three solutions using cube roots of 1 here. Cardano's methods were not fundamentally different from what you saw with the quadratic formula, just much more complicated.

The results for quartic equations (degree 4) are even more involved. However all of these formulas are of the same general type in that they only involve ordinary arithmetic operations and also radicals—square roots, cube roots, and degree 4 roots for the quartic. The standard term for a formula of this type is a "solution by radicals." What we've seen is that all quadratic, cubic, and quartic equations have

solutions by radicals. Notice that solutions by radicals can involve not only roots from the original a, b, c, and d values (as in the quadratic formula), but also roots of roots (as in the cubic equation).

The search for an analogous solution for equations of degree 5 and above lasted 250 more years. Finally in 1824 Abel proved that there could be no such general formula for equations of degree 5 and above. However the problem was only really understood in 1830 when Galois came up with a new way of looking at the subject that explained exactly which equations could be solved by radicals and why. Galois' paper was one of the most remarkable events in the history of mathematics, because it was a complete explanation of a fundamental problem and because its methods has formed the basis for work in many other areas of mathematics in the years since.

As noted Galois' solution was a matter of insight--a new way of looking at the problem that made everything fall into place—and this chapter is intended to provide enough detail to appreciate it. We don't prove everything, but try to show how his approach made this apparently strange result (why 2, 3, and 4 but not 5?) completely clear. As for Galois himself, his paper was unread or unappreciated until the 1850's, long after he was dead in a duel before the age of 21.

A closer look at the problem

Our goal is to understand which equations can and can't be solved by radicals. So we need to think a little more about what a solution by radicals really means.

We start with the basics. In any particular case, solution by radicals means you find the solutions by plugging into a formula involving ordinary operations of arithmetic plus radicals (square roots, cube roots, etc.). You can think of this as a generalization of the quadratic formula in two ways:

1. We allow not just square roots, but roots of arbitrary degree.

2. We also allow roots of roots, as noted for the cubic formula mentioned at the start.

To go farther we need a little more a little more terminology, starting with a couple of probably familiar definitions:

- The "integers" are the whole numbers both positive and negative, i.e. … -3, -2, -1, 0, 1, 2, 3 …

- The "rational numbers" are numbers that can be expressed as fractions of integers. That includes the integers themselves, since you any integer n is the same thing as n/1. So rational numbers look like look like -1/2, 0, 4, 2/3, etc. Standard terminology uses the bold symbol **Q** (for quotient) to denote the rational numbers. The rational numbers form a "field", that is to say a set of numbers you can add, subtract, multiply, and divide.

All of the coefficients of our polynomial equations and formulas are assumed to be rational numbers. If we're going to have a general formula, it has to work for polynomials with rational coefficients.

Galois' way of thinking makes solution by radicals a question about numbers: What kinds of numbers can you get from a formula for solution by radicals? Can those numbers solve everything?

As background for that approach, we first need to review the most basic solution by radicals, the quadratic formula.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

What the quadratic formula says is that even if a, b, and c are all rational numbers, we will need to add new numbers—square roots—to the rational numbers if we want to solve quadratic equations where $b^2 - 4ac$ isn't already a square.

We can see this in the specific case of the equation 2 x^2 + 3 x – 4 = 0. Plugging the values, 2, 3, and -4 into the quadratic formula gives

$$x = \frac{-3 \pm \sqrt{3^2 - 4\,x\,2\,x\,(-4)}}{2\,x\,2} = \frac{-3 \pm \sqrt{41}}{4}$$

The non-rational number we need to add is $\sqrt{41}$. And with that we can get both solutions $\frac{-3 + \sqrt{41}}{4}$ and $\frac{-3 - \sqrt{41}}{4}$.

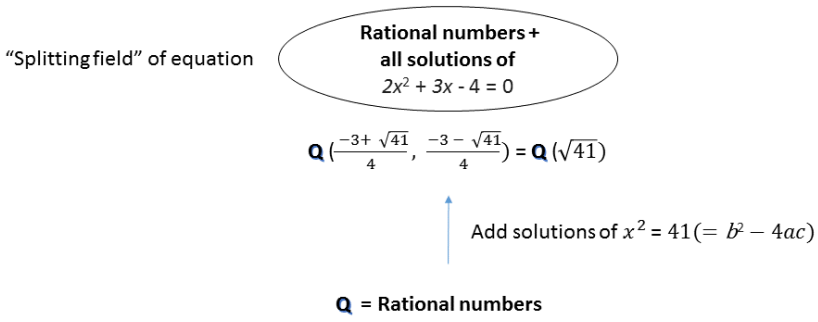
We need some additional notation to make this easier to generalize.

First we use the notation **Q** ($\sqrt{41}$) to mean the field you get by starting with the rational numbers and throwing in the new number $\sqrt{41}$. This is just like the way we got the complex numbers by adding $i = \sqrt{-1}$ to the reals.

Next we use the term "splitting field" for the equation 2 x^2 + 3 x – 4 = 0 to mean the field you get by throwing in all the solutions to 2 x^2 + 3 x – 4 = 0 to the field **Q** . By the quadratic formula that is **Q** ($\frac{-3+ \sqrt{41}}{4}$, $\frac{-3 - \sqrt{41}}{4}$). However **Q** ($\sqrt{41}$) includes every new number you can build with rational numbers and $\sqrt{41}$, which includes $\frac{-3- \sqrt{41}}{4}$ and $\frac{-3+ \sqrt{41}}{4}$. So the "splitting field" of 2 x^2 + 3 x – 4 = 0 is just **Q**($\sqrt{41}$).

Figure 1A shows this in graphical form. At the top we see the "splitting field" for the equation 2 x^2 + 3 x – 4 = 0, which is the same as **Q** ($\sqrt{41}$).

Figure 1A: Solving a Quadratic Equation



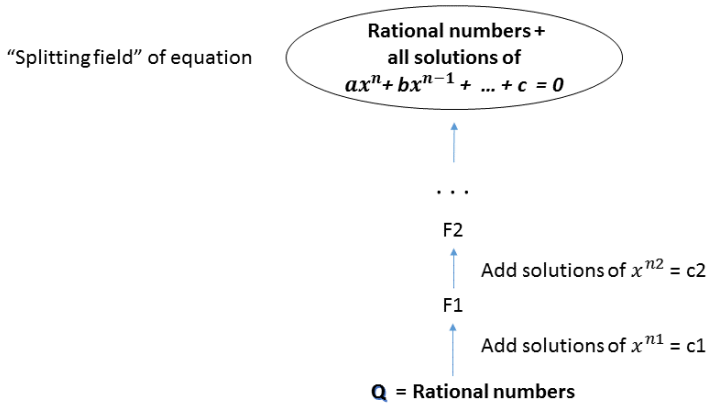
The quadratic formula says the same is true in general. For every quadratic equation $ax^2 + bx + c = 0$, the splitting field is $\mathbf{Q}(\sqrt{b^2 - 4ac})$.

We can take that as a definition of solution by radicals: for quadratic equations, solution by radicals means that we reach the splitting field for any quadratic equation $ax^2 + bx + c = 0$ by adding in the solutions of $x^2 = b^2 - 4ac$.

We can use these same concepts to say exactly what we mean by a solution by radicals in general.

To be specific, suppose we are given an equation $ax^n + bx^{n-1} + \ldots + c = 0$. We can talk about its "splitting field" which is just $\mathbf{Q}$ with all solutions added in. However, because the solutions are built up by evaluating radicals in the formula, we get to the splitting field by adding in the radical values—as we did with the quadratic case. Figure 1B gives the general version of what we saw in Figure 1A.

Figure 1B: Solving an Equation by Radicals



To clarify the notation, in Figure 1B the fields F1 and F2 are just names for the fields we get by adding the indicated values, like **Q** $(\sqrt{41})$ in the quadratic case. F1 adds the solutions to the first radical equation x^{n1} = c1 to the base field **Q** ; F2 adds solutions to the second radical equation x^{n2} = c2 to the field F1. Notice that c2 is an element of F1, so it can be an expression involving one of the added roots—which is what happens with the roots of roots in the cubic formula. As you go up in the chart, the fields get bigger, as more radicals are added. The bigger fields are called "field extensions" of the smaller ones.

To be clear—thus far we haven't solved anything, but we do have a more precise way of thinking about the problem.

For an equation to be solvable by radicals, the splitting field for the equation (i.e. the field you get by adding all the solutions of the equation to the rational numbers **Q**) has to be something you can build up step-by-step using roots of equations x^n = c.

In other words, all the roots have to be numbers you can build up, step-by-step using radicals. Can we find all the roots of all equations that way?

Galois' approach

Galois recognized that the crux of the issue was that the solutions of the equations x^n = c are related to each other. As an example we consider a simple case, the equation x^p = 1 (where p is a prime number) over the rational numbers **Q**. In that case (see Appendix 1) it turns out that all the solutions are actually powers of each other, and any solution ≠ 1 will generate the whole set. So the fundamental issue for solution by radicals is to know when a succession of related solutions to equations of type x^n = c can or can't solve a general equation.

To get at that issue Galois came up with a uniform way to keep track of relations among solutions for the different field extensions in Figure 1B. Very simply, for any of the field extensions he took the set of new solutions that were added and defined the Galois group to be the permutations of the solutions consistent with the relations among them. That sounds confusing, but actually it's not.

To understand it, we'll look at an example—the solutions to $x^5 = 1$ added to the rational numbers **Q** . From the Appendix we know the solutions are the number 1 and the four primitive 5^{th} roots whose powers each generate all five solutions. The number 1 is an ordinary rational number, so we only have to care about the 4 primitive roots, call them a, b, c, d. The permutations mentioned for the Galois group would be relabelings of the four solutions. Can we take just any relabeling? Could we call them d, c, b, a?

The answer is maybe not, precisely because the relations between the solutions limit the relabelings that we can use. For example, if currently a^2 = b, we'll need the same relation to be true with the new (relabeled) "a" and new "b".

What makes this $x^5 = 1$ case particularly easy is that all four non-rational solutions are powers of any one of them. Since all the solutions are powers of each other, once we decide where "a" goes then we know where all the powers of "a" must go, and therefore we know everything about the relabeling. That means there are only 4 possible "good" relabelings: the identity which sends "a" to itself and thus changes nothing, and the three others that send "a" to b, c, or d with the remaining labeling in each case determined by the values of the powers of a.

As we noted, the relabelings just discussed are "permutations"—reorderings of the four objects a, b, c, d (see Appendix 2 for a review of permutations). What we have seen is that out of the 24 possible permutations of the 4 objects a, b, c, d, only 4 are consistent with the relations among the 5^{th} roots. So the Galois group for $x^5 = 1$ over the rational numbers **Q** has just those four elements. Since the logic we used for p = 5 actually applies for any prime number p, the Galois group for the p^{th} roots of 1 over **Q** has exactly p – 1 elements.

We need to say more about the term "group." In the first chapter we defined groups to be collections of elements with a single operation—including an identity and inverses for each element. The specific groups we talked about were made up of numbers mod m. The elements of a permutation group are different—they are reorderings of an underlying set of objects—but we can still think of the set of permutations as just another set of objects where the group operation here performs the two permutations one after another. The identity is the permutation that does nothing. The inverse of any permutation just undoes it. (One of the reasons the group concept has proved so important is that it applies to many types of objects that are not numbers.)

For the Galois group, the point is that taking only the "good" permutations (consistent with relations among solutions) still leaves us with a group. Two good permutations done back to back are a good permutation (because two consistent permutations are still consistent). Also undoing a good permutation is also good (because we didn't violate consistency). So for any equation, the good permutations are a subgroup of the overall permutation group. That's why we can refer to the Galois "group".

In what follows you'll see that to an amazing degree, this notion of Galois group delivers precisely what is needed to understand solvability by radicals.

The Galois group of a general equation can be difficult to calculate, but we won't actually have to calculate very many different Galois groups here. If you look at Figure 1B, we obviously need to understand the Galois groups associated with the equation x^n = c. And we will want to compare those to the Galois group of a given equation $ax^n + bx^{n-1} + \ldots + c = 0$ over the rational numbers. However, if we're going to have a general formula to solve any equation of degree n, then we need to be able to handle the case that the solutions are independent of each other, that is—where the Galois group of the equation is the full permutation group on n objects. (As one would expect, that is the general case—see Appendix 4.)

So we now have things setup to begin looking in detail at Figure 1B, but we haven't quite gotten to the punch line. That punch line is that the Galois group of an equation $ax^n + bx^{n-1} + \ldots + c = 0$ tells you everything there is to know about possible subfields in Figure 1B! In particular if we are talking about a general formula for solution by radicals, all you need to do is look at the structure of the full permutation groups on n objects to know whether a formula can or cannot exist.

A little group theory and the Galois theorem

In order to state Galois' main theorem, we first need a basic concept from group theory. As it turns out, this is just a generalization of the modular arithmetic from the first chapter.

Suppose we have a subgroup H inside a group G. For any element g of G we use the notation gH to mean the collection of elements you get my multiplying every element of H by g. The usual term for this is gH is a "coset" of H in G. We saw cosets like that in Appendix 2 of the first chapter, where we remarked that cosets are either distinct or identical. That is because any element in common between sH and tH will have $sh_1 = th_2$. That means $s = th_2h_1^{-1}$. So s is in t's coset, and the two cosets are the same.

We would like to be able to do group operations with cosets the way we did arithmetic with remainders for modular arithmetic. So we would like sH x tH to be the coset stH. The only is trick to this is that groups are not necessarily commutative (ab is not always = ba even for permutations), so we don't necessarily have tH = Ht, and that gets in the way. We solve that problem by simply excluding such cases by definition. We say

- N is a "normal" subgroup of G if gN = Ng for all elements g in G. (Note this doesn't say anything about G being commutative, it just says that as sets of elements gN and Ng are the same.)
- For a normal subgroup N of G we define the quotient group G/N to be the group whose elements are the cosets gN and the operation is sN x tN = stN .

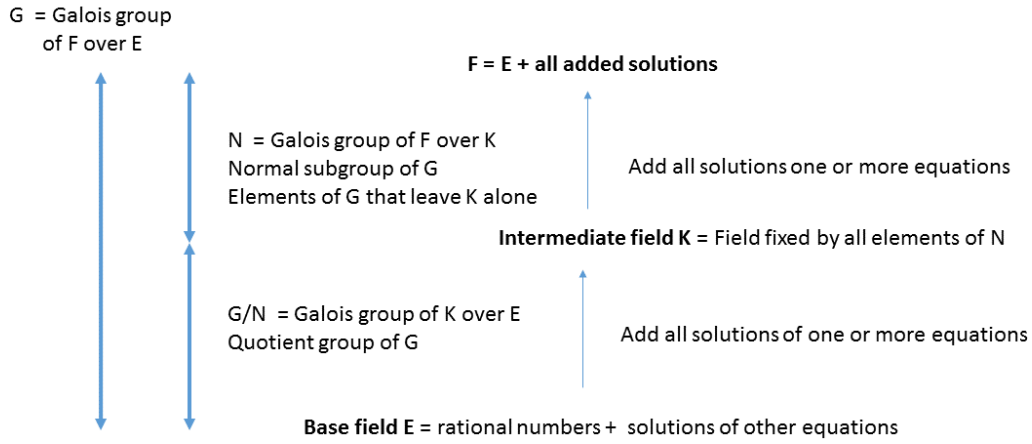
In case it's not obvious, we can be specific about the parallel with modular arithmetic. Take G to be the set of integers, positive and negative, with addition as the operation. Then take N to be the set of all multiples of 7, also under addition. Then G/N is just the additive group for arithmetic mod 7. Each coset corresponds to a remainder mod 7.

It should be noted that the group concept itself was rudimentary in Galois' time, and the notion of "normal" subgroup was defined by him for the purpose you'll see next.

We're now ready to look at what a Galois group tells us about possible field extensions. The picture is actually simple. Figure 2 tells the bottom line.

The basic picture is that we have a field extension F of a base field E with Galois group G. That extension is built up in two stages with an intermediate field K. As figure shows, we end up with an explicit correspondence between normal subgroups N of G and intermediate extensions K between F and E.

Figure 2: Galois Theorem:
Intermediate fields K correspond to Normal subgroups N of G



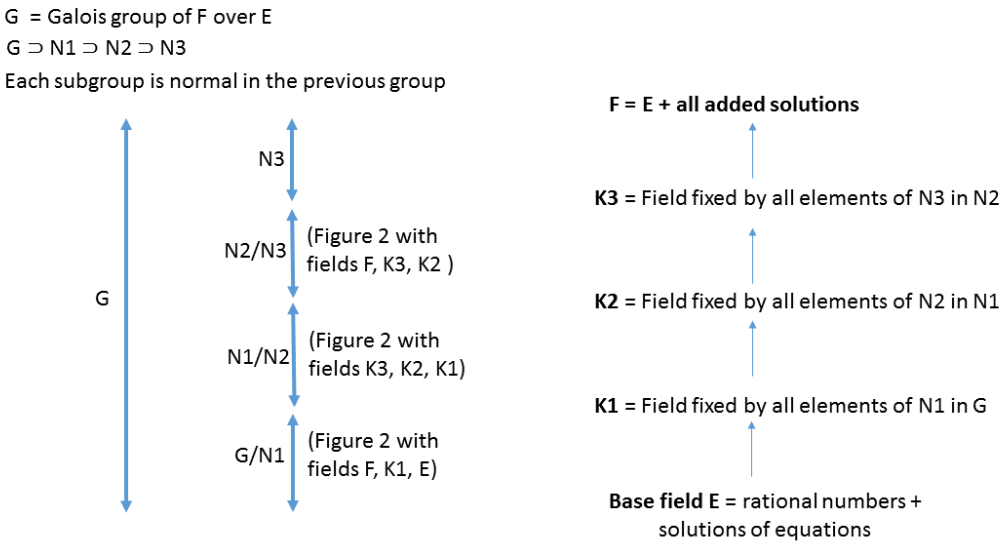
It takes a little effort to understand the details of that correspondence, so we're put that discussion in Appendix 3. For now, the important fact is that there is such a correspondence. The only way we can have such an intermediate field K between E and F is if we can find an appropriate normal subgroup N in G.

Multiple Intermediate Fields

As it turns out, the Galois correspondence gives even more than you might think, because the generality of Figure 2 means we can apply it not just to single intermediate fields but to configurations like Figure 1B. That again sounds abstract, but Figure 3 makes that concrete. We can be specific about what to look for in the Galois group G if there are to be multiple intermediate fields. If we're going to have a chain of increasing fields as in Figure 1B, they have to correspond to a chain of decreasing normal subgroups in G.

Figure 3 looks complicated, but actually there's not much to it. It's just Figure 2 repeated over and over again as you go down the chain of subgroups or up the set of field extensions. The left side of the figure shows exactly what the Galois groups are at each stage.

Figure 3: Galois Theorem with Multiple Intermediate Fields



Specifically, in Figure 3 we have a top field F and a bottom field E with three subfields K1, K2, and K3 in between. Those three subfields correspond to three subgroups N1, N2, and N3 of the Galois group G of F over E. N1, N2, and N3 are defined exactly as in Figure 2 using the correspondence between subfields and subgroups of G. Because those fields are contained in each other (F ⊃ K3 ⊃ K2 ⊃ K1), we have G ⊃ N1 ⊃ N2 ⊃ N3. Notice the ordering of the groups is reversed from the ordering of the fields, because the lowest field is farthest from the top, so it has the largest permutation group for solutions.

What is interesting about Figure 3 is that it gives a specific subgroup or quotient group of G for every individual field extension F ⊃ K3 ⊃ K2 ⊃ K1⊃ E (note the column of vertical arrows labeled by group names or quotient groups). We did that by applying Figure 2 to three different sets of fields, as indicated to the right of each quotient group.

We can summarize Figure 3 as follows. In order to have the field extensions F ⊃ K3 ⊃ K2 ⊃ K1⊃ E, the Galois group G of F over E must have subgroups G ⊃ N1 ⊃ N2 ⊃ N3 where

- Each subgroup is normal in the group above, so the quotient groups G/N1, N1/N2, and N2/N3 are well-defined.
- The quotient groups G/N1, N1/N2, and N2/N3 are the Galois groups of K1 over E, K2 over K1, and K3 over K2 respectively. (Note that the quotient groups just represent each pair of subgroups going down the inclusion list G ⊃ N1 ⊃ N2 ⊃ N3.)
- N3 is the Galois group of F over K3.

There is of course nothing special about having three intermediate subfields K1, K2, K3—we get the same set of results for any number of intermediate fields K.

With this we can now relate every incremental field extension in Figure1B to a corresponding subgroup or quotient group of the overall Galois group G. As a result, once we understand the Galois groups

corresponding to the equation x^n = c, we can know exactly what we would have to find in the Galois group G for an equation to be solvable by radicals.

Solution by Radicals

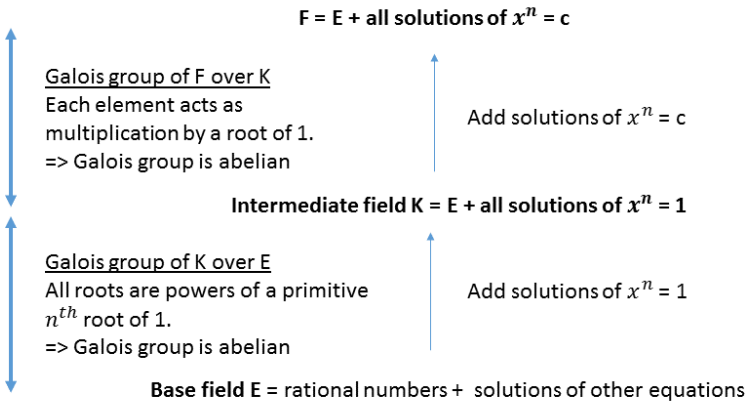
There are two parts to this section. First we complete our analysis of the Galois group for the equation x^n = c. That will tell us what the Galois groups look like for the field extensions given in Figure 1B. Then we will put the pieces together to see which equations can be solved by radicals.

The analysis of x^n = c turns out to be simpler than you might expect for the following reasons:

- The field we get by adding all solutions to x^n = c actually contains the solutions to x^n = 1. That is because the ratio of any two solutions of x^n = c is a solution to x^n = 1 (and we get all of them by the next comment).
- For the same reason if "a" is any solution of x^n = c, then you can get every solution of x^n = c by multiplying a by some solution of x^n = 1 (use the ratio of the new solution to the old one).

Figure 4 shows how this plays out. What the Figure shows is that you can build up the field extension corresponding to x^n = c in two stages: first throw in the roots of 1 and then add a root of c. If we do it that way, the Galois groups of the two stages are easy to describe.

Figure 4: Galois Groups for x^n = c



In Figure 4 the first extension is by n^{th} roots of 1. For this the analysis of x^n = 1 is actually very close to what we saw for x^p = 1, just a little more complicated to describe. The solutions can still be described explicitly as complex numbers, and you can still identify the primitive solutions whose powers generate all solutions. (To be explicit about this—following Appendix 1—the n^{th} roots of 1 are $e^{2i\pi k/n}$ for, k = 0,

1, 2,…,n-1. A root will be primitive if k has no prime factors in common with n. Notice that the primitive n^{th} roots of 1 correspond to the members of the multiplicative group mod n of chapter 1.)

As before, the elements of the Galois group for this equation are the identity (which changes nothing) and the other roots that the chosen primitive root "a" might go to. Since the operations of the Galois group amount to adding exponents for the generator, the order of the operations doesn't matter. Standard terminology for this is to say that the Galois group is "abelian" (i.e. for any elements a and b of the group ab = ba).

The analysis of the second extension F over K in Figure 4 is even simpler. The field F adds a root of x^n = c to the field K which already contains all the n^{th} roots of 1. To get the Galois group of F over K we use the earlier remark that every solution of x^n = c can be obtained from a given one by multiplying by an n^{th} root of 1. That means that every element of the Galois group of F over K acts on each solution by multiplying by an n^{th} root of 1. Since ordinary multiplication is obviously independent of order (2 x 3 = 3 x 2 = 6), we can conclude that the Galois group of F over K is also "abelian".

We can now put the pieces of our argument together. Figure 5 is a reworking of Figure 1B in light of what we have just learned. Each of the field extensions F1, F2 has been divided into two parts based on what we know about the equation x^n = c. As we concluded, each one of these extensions has an abelian Galois group.

Figure 5: Solving an Equation by Radicals Revisited

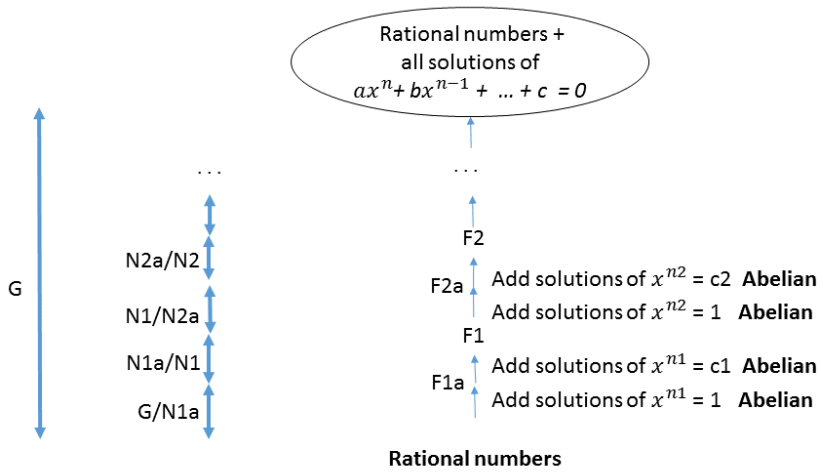


Figure 5 also shows the correspondence of field extensions and subgroups as originally presented in Figure 3. From Figure 5 we get the following for G = Galois group of F over the rational numbers:

- The field extensions F ⊃ ⋯ . ⊃ F2 ⊃ F2a ⊃ F1 ⊃ F1a ⊃ E correspond to subgroups G ⊃ N1a ⊃ N1 ⊃ N2a ⊃ N2 ⊃ …

- Each subgroup is normal in the group above it, so the quotient groups of consecutive groups are well-defined.
- The quotient groups G/N1a, N1/N1a, N2a/N1 … are the Galois groups of the individual extensions starting from the bottom of Figure 5—that is for F1a over the rationals, F1 over F1a, and F2 over F1a and so forth. See the vertical arrows on the left side of the Figure.
- The final (smallest) group in the series is the Galois group of F over the highest field
- Since we know that the Galois groups of all the individual extensions are abelian, the same must be true for all of the quotient groups and also for the final group in the series.

What this boils down to is that for a given polynomial equation to be solvable by radicals we need its Galois group G to have a decomposition by subgroups G ⊃ N1 ⊃ N2 ⊃ … ⊃ NX where each subgroup is normal in the previous with quotient group abelian, and the final group NX is also abelian. So either G is abelian itself or it has a decomposition by such subgroups. Following Galois, such a group is called "solvable".

As was noted before, a general formula for solvability by radicals (such as the quadratic formula) has to cover the case where there are no relations between the solutions—i.e. where the Galois group is the full permutation group S_n on n elements (see Appendix 2). So we can have a general formula for solutions of all equations of degree n only if the permutation group S_n is a solvable group. The results from Appendix 2 can be summarized as follows:

- For n = 2, 3, 4 the permutation groups on n elements can be explicitly calculated. The n = 2 case is trivial as the permutation group has only two elements; for n = 3 the group is non-abelian but has an abelian normal subgroup with abelian quotient; for n = 4 the group is again solvable, this time based on a decomposition with two subgroups.

- For n = 5 the permutation group has 120 elements and is non-abelian. It has exactly one normal subgroup—the even permutations—with 60 elements. Since the quotient group has only two elements, that quotient is abelian. However the normal subgroup itself is non-abelian and has no normal subgroup, so S_5 is not solvable.

- For n = 6 and above the situation is the same as with n = 5, as can be proved by induction starting with n = 5.

And that—for 2, 3, 4 but not 5 or more—is that.[†]

[†]With a few last details in Appendix 4.

We can, however, go one step farther. Already for Galois the full statement of the result was that an equation is solvable by radicals if and only if its Galois group is a solvable group.

Thus far we've done the "only if" side. As it turns out, the "if" side of the argument falls out (with work) from the same analysis, because the decomposition by subgroups can be turned into an explicit construction of a solution by radicals. The procedure is related to our analysis of x^n = c. For each abelian group in the decomposition of a solvable group you add-in appropriate roots of 1 and then prove there are appropriate "c" values to finish the job. The quadratic, cubic, and quartic formulas can all be derived in this way. For details see Galois Theory by Ian Stewart, chapter 15. With that, the argument for solution by radicals is complete.

It remains only to talk about the implications of this work.

Not only was this a complete solution ("if and only if") to the very basic question of solvability by radicals, but it was a new conceptual way of thinking about the problem. It's not often that one comes to the answer of a 200-year-old problem and understands exactly why it is true.

That new way of thinking has had an enormous influence on subsequent mathematics. Most obviously the notion of Galois group, with the association of subgroups with subfields, has become part of the mathematical vocabulary for all kinds of topics. However even more fundamentally, Galois work was one of the first examples of the power of abstract groups, and as we've already seen—groups turn up everywhere.

One way to think of mathematics in general is that it is about what has to happen just because of the way numbers work. Galois understood an important piece of that, and settling the question of solution by radicals was a significant byproduct.

Interestingly enough our next chapter--on Fermat's Last Theorem--is a fitting sequel in two different ways:

- It is another case of fundamental understanding driving a significant byproduct.
- The notion of Galois group is a surprisingly important piece of the puzzle.

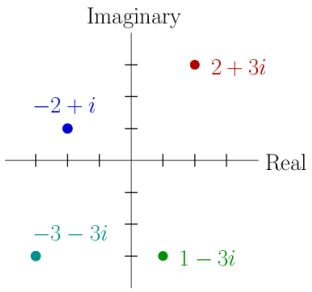
We're now ready to look at one of the most important pieces of 20th century mathematics ….

Appendix 1: Complex numbers and roots of one

This appendix is a very brief summary of some properties of the complex number plane. More complete discussions are available many places, including the Khan Academy.

Just as ordinary real numbers are visualized as a line, complex numbers a + bi (i = $\sqrt{-1}$) are visualized as points in a plane, with independent real parameters a and b. The value a is called the real part, and bi is the imaginary part as in Figure A1.

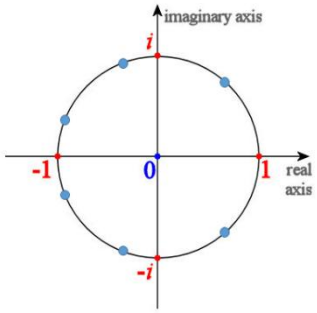
Figure A1: The Complex Plane



As we noted in the last chapter, the Fundamental Theorem of Algebra (originally proved by Gauss) says that every root of every polynomial can be expressed as a complex number a + bi with real numbers a and b. This is obvious in the quadratic formula, for example, since the radical $\sqrt{b^2-4ac}$ is equal to a real number if b^2-4ac is positive or else equal to i$\sqrt{4ac-b^2}$ otherwise.

Roots of 1 are represented in a particularly nice way in the complex plane, as they sit as division points on the unit circle. For example, the 7th roots of 1 start with value 1 on the x-axis and then go around the circle in equal segments with the values cos(2πk/7) + i sin(2πk/7) for k = 1,2,3,4,5,6 as in Figure A2.

Figure A2: Unit Circle and 7th Roots of 1



There is formula from Calculus that shows this in a particularly nice way. That remarkable formula, due to Euler, says that e^{ix} = cos x + i sin x. (Here e is Euler's number, approximately 2.718--the "natural" exponential, and x is in radian units. Hence the amazing result $e^{i\pi}$ = -1.)

We're not going to discuss the formula in any detail here, but what is nice about it is that it converts the expression cos x + i sin x into an exponential. That makes our example of the 7th roots of 1 completely obvious, since for k = 1,2,3,4,5,6 we have

$$\cos\left(\frac{2\pi k}{7}\right) + i\sin\left(\frac{2\pi k}{7}\right) = e^{i2\pi k/7},$$

so $(\cos\left(\frac{2\pi k}{7}\right) + i\sin\left(\frac{2\pi k}{7}\right))^7 = (e^{\frac{i2\pi k}{7}})^7 = e^{i2\pi k}$ = cos(0) + i sin(0) = 1. (If you remember Trigonometry, just about everything with trig identities comes down to Euler's formula! For example you can get the addition formulas for the sine and cosine by applying Euler's formula to each side of the usual rule of exponents e^a x e^b = e^{a+b} . Remember that saying two complex numbers are equal means that both the real and imaginary parts are equal.)

Obviously the same holds for any nth roots of 1. In that case the primitive roots are the ones with k relatively prime to n. That's because the other ones have some factor f in common with n, and therefore the n/f power will be equal to 1—so you can't possibly get all the roots. For the primitive ones only the nth power is 1, so the powers of any of them do generate all the others.

Appendix 2: Permutations and Permutation Groups

The intent of this section is to summarize the basics about permutations and permutations groups. There are of course many references for this topic, including Wikipedia and virtually any course book in algebra.

To start, a permutation is a reordering a set of objects. Suppose we have four objects. Then a permutation of those four objects says what happens to the object in each of four numbered slots. For example the permutation notation $\binom{1\,2\,3\,4}{2\,1\,3\,4}$ indicates that the first two objects should be switched in position and the other two objects left alone. Since you can assign a first object to any of the four slots, the second to any of the remaining three, and so forth, there are 4 x 3 x 2 x1 = 24 possible permutations of four objects. For n objects the number of possible permutations is n x (n-1) x …x1 = n!.

The $\binom{1\,2\,3\,4}{2\,1\,3\,4}$ notation is perfectly general, but there is a more descriptive way to think about permutations. That alternative way uses the notion of cycles. A cycle is best defined by an example: the cycle (123) moves the first object to the second slot, the second object to the third slot, and the third object back to the first. Cycle notation specifies just the cycles included in the permutation; objects that stay put are omitted. For example the previous permutation $\binom{1\,2\,3\,4}{2\,1\,3\,4}$ is just (12) in cycle notation.

For the permutations of 4 objects the 24 elements as expressed in cycle notation are

(12), (13), (14), (23), (24), (34)

(123), (124), (132), (134), (142), (143), (234), (243)

(12)(34), (13)(24), (14)(23)

(1234), (1243), (1324), (1342), (1423), (1432)

No change = the identity permutation

Permutations such as (12) that only switch pairs of elements are called “transpositions”. Every permutation can be built up as a sequence of transpositions. While this can be done in different ways, the number of such transpositions will always be either even or odd for a given permutation. So any permutation is called either even or odd based on the number of transposition in the sequence to build it.

Permutations of n objects form a group where the group operation is defined as performing the two operations in sequence. In this group the identity is the permutation that does nothing. For any permutation, the inverse is the permutation that undoes it. Standard notation for the full permutation group on n elements is S_n called the “symmetric group on n elements”. For any n the <u>even</u> permutations form a normal subgroup of index 2 (i.e. half the elements) denoted A_n and called the “alternating group”.

Interestingly, most permutation groups are “non-abelian” (or “non-commutative”—a x b is not always = b x a). $S_2$is trivial with only 2 elements, but already S_3 is non-abelian where (12) x (13) = (123) but (13) x (12) = (132).

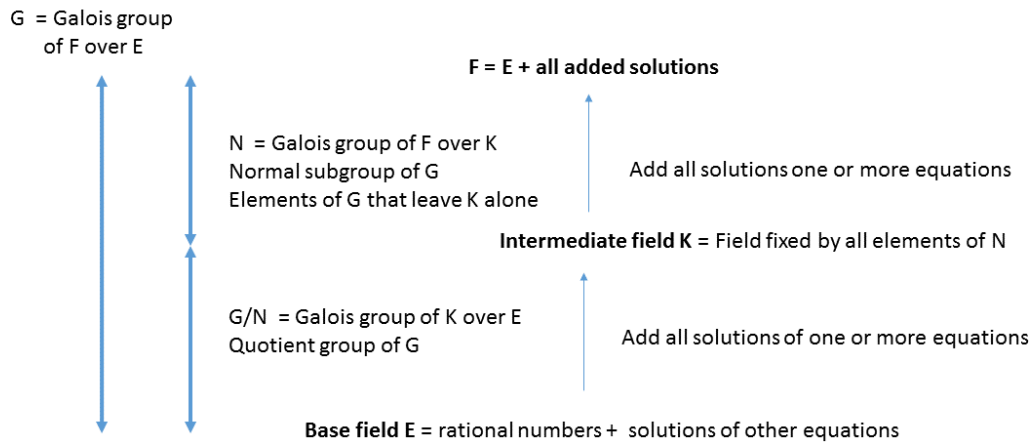
Of particular interest for this chapter is whether S_n is a "solvable" group (see the definition in the text).

- S_2 is trivially solvable, because it has only two elements and is therefore abelian.
- S_3 has A_3 as a normal abelian subgroup and is therefore solvable.
- S_4 has A_4as a normal subgroup with abelian quotient, but A_4 itself is not abelian. However A_4 has an abelian normal subgroup with abelian quotient—this is the subgroup with the three double transpositions (e.g. (12)(34)) plus the identity. Therefore solvable.
- For S_5 and above, S_n has A_n as a normal subgroup, but A_n is simple (no normal subgroups) and non-abelian.

Appendix 3: Understanding the Galois Correspondence

The goal here is to understand the Galois correspondence from Figure 2.

Figure 2: Galois Theorem:
Intermediate fields K correspond to Normal subgroups N of G



To do that, we first have to think about the Galois group is a slightly different way. Thus far we talked about the elements of the Galois group G as permuting the solutions we added to the base field E to get the bigger field F. However, you can also think of the elements of G as acting on the bigger field F itself. That's because every number in F is made out of the added solutions with numbers in the base field E as coefficients. So for any element in F we can let the elements of G act on the added solutions as usual and leave everything else alone. This is trivial, but needs an example to make it clear:

Consider the case where F is the field you get by adding solutions to x^2 = 3 to the rational numbers **Q**. The solutions to the equation are $+\sqrt{3}$ and $-\sqrt{3}$, and the numbers in the field F look like a + b$\sqrt{3}$, where a and b are ordinary rational numbers.

The Galois group has two elements:

1. the identity permutation that does nothing, and

2. the permutation that switches $+\sqrt{3}$ and $-\sqrt{3}$.

We can now be very explicit about the action of the Galois group G on the numbers in this field F. As we said, we let the elements of Galois group act on the new values (i.e. $\pm\sqrt{3}$) and leave everything else alone. The identity element changes nothing, so there is nothing to say. The other element shows what happens. Specifically a + b$\sqrt{3}$ becomes a - b$\sqrt{3}$ and vice versa. That's all there is to it. The general case works the same way—the elements of G work on the added solutions and everything else stays the same.

Based on this idea of elements of G acting on numbers in F, we can explain the terminology in Figure 2. If we pick some number f in the field F, we can talk about the elements of the Galois group that leave it unchanged. Obviously if f is really in the base field E, then every single element in G leaves it alone—just by definition. In general if an element g of the Galois group leaves f unchanged, we say g "fixes" f. (An element f could for example be a sum of a set of elements permuted by g. In that case it would be fixed by g but not in E.) We can use the same terminology for the whole intermediate field K: we say g "fixes" K if it leaves every single number k in K unchanged.

In Figure 2 the correspondence between normal subgroups N of G and intermediate extensions K between F and E is based on this notion of "fixed" intermediate fields K. If we start with an extension K, then the set of all elements g that "fix" K turns out to be a normal subgroup—call it N. On the other hand if we start with a normal subgroup N, then the corresponding field K is the set of numbers in F "fixed" by every member of N. The theorem says with that correspondence, N is the Galois group of F over K and G/N is the Galois group of K over E.

This Galois correspondence is a big step, because it gives us direct relationship between subgroups of a known Galois group and possible subfields K in Figure 1B. It is also actually intuitive, as we describe next.

To see where the correspondence comes from, let's imagine to get K from E you add in a first set of new roots R1 and similarly to get F from K you add in a second set of new roots R2.

The first thing to notice is that all the roots of R2 are in the field F, and that both G and the Galois group of F over K contain the "good" permutations of those same roots. So we can identify the Galois group of F over K with the elements in G that exclusively permute the roots in R2. There is even a simple way to say which elements in G they are. Since the roots R2 are precisely what gets added to go up from K to F, the Galois group of F over K is made of up the elements of G that leave K alone.

Let N denote that collection of elements of G that leave K alone. This is obviously a subgroup (a combination of permutations that don't affect K also doesn't affect K). However it is almost equally obviously a "normal" subgroup. That is because if g is some element of G not in N, then any element in gNg^{-1} doesn't affect K (since N does nothing, any effect of g gets undone), so gNg^{-1} = N or gN = Ng.

The next question is what we can say about the Galois group of K over the base field E. We can't necessarily say that it involves a subgroup of G, because it may be that the elements of G that permute roots in R1 also affect roots in R2. (This can happen if the roots in R2 are actually roots of elements of R1, as in the cubic formula). However since N is normal, we can "divide out" the portion of G that we just associated with F over K and get the Galois group of K over E! Specifically since N leaves K alone, all the elements in the coset gN perform the same permutation of the roots in R1. Further the cosets must represent distinct permutations (or something else would fix K), and we must get all "good" permutations in this way (or G would not be complete). So the Galois group of K over N is the quotient group G/N!

So to get from K to N, you find N as the subgroup of G that fixes K. This N is the Galois group of F over K and turns out to be normal in G. Further the quotient group G/N is the Galois group of K over E.

In the opposite direction, if you start with a normal subgroup N in G, you can find K as the biggest subfield of F that is fixed by everything in N. And the Galois groups are again N and G/N.

Appendix 4: Last Details

We end the chapter with two last details worth discussing.

First we noted (without proof) that if we're going to have a general formula to solve any equation of degree n, then we need to be able to handle the case that the solutions are independent of each other, that is—where the Galois group of the equation is the full permutation group on n objects.

It is true, as we mentioned, that is the general case. However the proofs are non-trivial and we can actually get by with far less. In fact to prove there is no general formula for an n all we really have to care about is finding one single polynomial of degree 5 that is not solvable by radicals! That's because if p5(x) is our non-solvable polynomial for degree 5 then $p5(x)(x-1)^{n-5}$ is a polynomial of degree n that is also non-solvable by radicals.

For this, the Ian Stewart book mentioned earlier gives p5(x) = x^5 - 6x + 3 as an example. The graph of this polynomial crosses the x axis three times, so it has 3 real roots and one pair of complex conjugate roots. Swapping the complex conjugates gives an element of the Galois group of degree 2. Using arguments from group theory, he shows there is also a group element of degree 5. And in the specific case of symmetric group S_5 those two elements generate the whole group, so we're done.

For the harder problem of finding nth degree polynomials whose Galois group is all of S_n this comes down to the so-called "Hilbert irreducibility theorem," which shows that there are infinitely many such cases—but without constructing any. Beyond that, there are lower bounds on the fraction of nth degree polynomials that have Galois group equal to S_n. (Interestingly, the Hilbert irreducibly theorem also turns out to play a role in the proof of Fermat's Last Theorem of Chapter 4, although the details are beyond the scope here.)

Second (and more technically) we have assumed for simplicity that the splitting field of the polynomial is exactly equal to the field built up by radicals. In fact for a solution by radicals we only need the splitting field to be contained in a field built by radicals. It doesn't matter if there is something extra, as long as we have what we need. So we have to allow for the case that the splitting field is contained in—but smaller than—the field built by radicals.

That gets resolved by applying the Galois theorem from Figure 2 once again, this time to the fields:

F = field built up by radicals ⊃ K = splitting field for the equation ⊃ E = rational numbers

Following Figure 2, we let G = Galois group of F over E and N = normal subgroup of G corresponding to K, then we get Galois group of the solution field K over E is the quotient group G/N. So this is now just a question for group theory—what can you say about a quotient group of a solvable group? In fact every subgroup and quotient group of a solvable group is solvable (because subgroups and quotient groups of an abelian group are abelian), so we end this by saying:

Solvable by radicals => Galois group of radicals field is solvable => Galois group of polynomial is solvable.

Chapter 4. Pythagorean Triangles, Conic Sections, and Fermat's Last Theorem

Background

Fermat's Last Theorem had a more than 300-year history as perhaps the most famous unsolved problem in mathematics. The story around its solution is both dramatic and mathematically extraordinary. This chapter tries to give a flavor of both.

The Theorem is simple to state: the equation $x^n + y^n = z^n$ has no solutions with whole numbers x, y, and z for any n ≥ 3. It is "last" in the sense that all other assertions Fermat made were proven centuries ago, whereas this one remained a puzzle. Fermat himself claimed he knew how to prove it, but his "method of descent" is unlikely to have gone much beyond n = 4.

For n = 2 we have all seen simple solutions with Pythagorean triangles such as $3^2 + 4^2 = 5^2$ or $5^2 + 12^2 = 13^2$. We begin with that case, as there is a general formula that gives all the (infinitely many) solutions in integers. The argument also gives a flavor of what it has traditionally meant to solve equations in integers.

After that we move on to the bigger story. The method of proof for Fermat's Last Theorem is actually only one step away from the familiar conic sections—parabolas, ellipses, hyperbolas—in that conic sections describe the solutions of equations of degree two, and the Fermat's Last Theorem finally succumbed to the mathematics of equations of degree three. Since the exponents in Fermat's Last Theorem can be arbitrarily large, the connection to equations of degree three was unexpected. And the proof was a surprise byproduct of a major piece of seemingly-unrelated work.

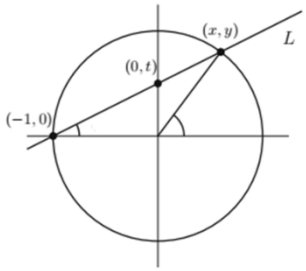
Pythagorean triangles

We begin with Pythagorean triangles (following [11])

The first thing to note is that Pythagorean triangles $x^2 + y^2 = z^2$ correspond to rational points on a circle of radius one. That is just a matter of taking $x^2 + y^2 = z^2$ to $(x/z)^2 + (y/z)^2 = 1$. Since real triangles will have x, y, and z positive, we only care about the top-right quadrant of the circle.

We now find the rational points using a very specific geometric argument based on the picture shown in Figure 1.

Figure 1: Rational Points on a Circle



In the figure there is a line drawn from a left hand point where the circle meets the x-axis to an arbitrary point (x,y) on the top-right quadrant of the circle. The line crosses the y-axis at what turns out to be a convenient point (0,t). We use the figure to express both x and y in the top-right quadrant in terms of the value t.

Since the point (x,y) is on both the line and the circle, we can get two equations for it. First since the line goes through both (x,y) and (0,t), the values y/(x + 1) and t/1 are equal because both are equal to the slope of the line. So we have y/(x +1) = t or y = t(x + 1). Then since x and y satisfy the equation of the circle, we have $x^2 + y^2 = 1$. Putting the two together we have 1 - $x^2 = y^2 = t^2(x + 1)^2$. Now since the left side 1 - $x^2 = (1 + x)(1 - x)$, we can actually remove a factor of (1 + x) from both sides (x= -1 corresponds to the trivial solution point (-1, 0) where the circle and the line meet on the x-axis). And we are left with (1 – x) = t^2(x + 1), which lets us rather easily solve for x in terms of t. The result is

$$x = \frac{1-t^2}{1+t^2}$$

Plugging in this value of x into the equation y = t(x+1) then gives (using 1 = $\frac{1+t^2}{1+t^2}$)

$$y = \frac{2t}{1+t^2}$$

Since these values for x and y will be rational numbers if and only if t is a rational number, we've actually (with very little work) found all rational points on the top-right quadrant of the circle. They are the points you get from our formulas for x and y as t takes rational values from 0 to 1.

The remaining work is to turn those rational solutions into integers X, Y, and Z. An important point is that any solution set X, Y, Z for the equation $x^2 + y^2 = z^2$can be turned into another solution set just by multiplying all three values by the same number. So to find all solutions, you have to find the primitive ones, that is the ones with no common factor.

The first thing to notice is that with a primitive solution $X^2 + Y^2 = Z^2$, exactly one of X and Y will be odd. It's obvious they can't both be even, since 2 would divide all three numbers. They also can't both be odd, because the square of an odd number is = 1 (mod 4). (An odd number = 2n + 1, and $(2n + 1)^2$ =

$4n^2$ + 2 x 2n + 1.) So the sum of two odd squares would be = 2 (mod 4), but Z^2 as an odd square would be = 1 (mod 4). For now we will assume we have a solution set X, Y, Z with X odd.

With that in mind, we now go back and see how to use our formulas for rational points on a circle to generate primitive solutions X, Y, Z. What we know is that for our given primitive solution X, Y, Z, X/Z and Y/Z are rational points on the unit circle, so that (from our formula for rational points on a circle)

$$X/Z = \frac{1-t^2}{1+t^2} \text{ and } Y/Z = \frac{2t}{1+t^2} \text{ for some rational number t.}$$

Since t is a rational number we can write it as t = m/n, where m and n have no common factors. When we plug that value for t we get

$$X/Z = \frac{n^2-m^2}{n^2+m^2} \text{ and } Y/Z = \frac{2nm}{n^2+m^2}.$$

We'd like to say that X = $n^2 - m^2$, Y = 2nm, and Z = $n^2 + m^2$, but we can't do that until we check that there is no common factor among those values.

Any common factor of X and Z divides their sum $2n^2$ and their difference $2m^2$. Since m and n have no common factors, that means the only possible common fact is the 2. If there is a common factor of 2 then $n^2 - m^2$ must be divisible by 2, but not by 4 since the quotient by 2 must be the odd number X. That means $n^2 - m^2$ = 2 (mod 4). But the difference of two squares can only be = 0 or 1 (mod 4), so 2 is also out and there is no common factor. Therefore the formula: X = $n^2 - m^2$, Y = 2nm, and Z = $n^2 + m^2$ will get every primitive solution of $X^2 + Y^2 = Z^2$ with X odd.

That formula won't get the primitive solutions with Y odd, because the value for Y is always even. So to complete the list we just switch X and Y to get the formula: X = $2nm$, Y = $n^2 - m^2$, and Z = $n^2 + m^2$. That gets all the primitive solution sets with Y odd by exactly that same logic as for X odd.

With those two formulas we generate all the primitive Pythagorean triples. All the non-primitive ones can be generated from the primitive ones by multiplying with common factors.

Starting Fermat's Last Theorem

We're now ready to move on to Fermat's Last Theorem

As a preliminary remark we note that in the years since Fermat's time there have a been a great many partial results for special values of n (see [2]). For our discussion here we can assume that the special cases of n =3 and 4 have already been done. With that, it is enough to prove the theorem for odd prime exponents p ≥ 5 That is because if our exponent n = n_1 x n_2, then $x^n + y^n = z^n$ can be rewritten as $(x^{n1})^{n2} + (y^{n1})^{n2} = (z^{n1})^{n2}$. Then since any n ≥ 5 will have a factor of 3, 4, or a prime p ≥ 5, any counterexample to Fermat's Last Theorem will produce a counterexample with n = n2 = 3, 4, or a prime p ≥ 5. So we're now looking for integer solutions to $x^p + y^p = z^p$ for primes p ≥ 5.

To get there we need to go beyond the degree 2 case of conic sections and Pythagorean triangles to equations of degree 3—so-called "elliptic curves". As we noted at the beginning, there was nothing obvious about the connection between elliptic curves and Fermat's Last Theorem. That connection was made using a major piece of recent mathematics called the Modularity Theorem—which classifies all

elliptic curves with their associated properties. The logic of the proof is that a counterexample to Fermat's Last Theorem leads to a strange elliptic curve, which in turn cannot exist by the Modularity Theorem. Beyond the connection to Fermat's Last Theorem, the Modularity Theorem is itself an amazing result, as we hope to show when we get there.

Historically, the Modularity Theorem goes back to conjectures first formulated in the 1950's, which became mainstream by 1960's. By the 1970's it was one of the major activities in mathematics, involving a highly-international cast of characters. The connection to Fermat's Last Theorem was first conjectured in 1985 and then established in 1986. The proof of the Modularity Theorem itself (in the cases needed for Fermat's last theorem) was announced by Andrew Wiles in 1993, but a gap was found in the argument. After more than a year of frantic effort the gap was finally closed by Wiles assisted by his former student Richard Taylor—and that completed the proof of Fermat's Last Theorem. Wiles' methods were extended to cover all cases for the full Modularity Theorem in 2001.

To cover all this territory, the remainder of this chapter is organized as follows:

- We first spend some time with elliptic curves, their properties, and the techniques people use to investigate them.
- We then go on to look at approaches used to classify elliptic curves. Here there is an interesting contrast between classical approaches and the newer direction that ultimately led to the proof of Fermat's Last Theorem.
- We can then talk about the Modularity Theorem in enough detail to understand its implications.
- Finally we cover the connection between the Modularity Theorem and Fermat's Last Theorem, which completes the proof.

Note that this logical order is different than the historical order, as the connection between the Modularity Theorem and Fermat's Last Theorem was established before the Modularity Theorem was actually proved. Wiles himself said that the connection to Fermat's Last Theorem was a major motivator for his intense work on the Modularity Theorem.

Elliptic Curves

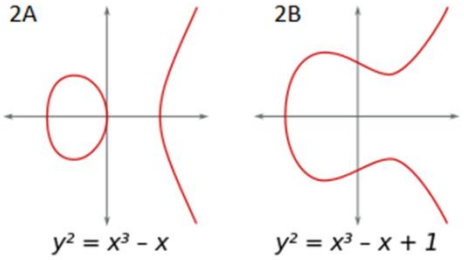
As mentioned earlier, "elliptic curves" are solutions to equations of degree 3. (There is an indirect connection to ellipses, but the term is mostly historical.) The general equation of degree 3 looks rather complicated:

$Ax^3 + Bx^2y + Cxy^2 + Dy^3 + Ex^2 + Fxy + Gy^2 + Hx + Iy + J = 0.$

However by standard changes of variable we can put the equations is in the following simpler form:

$y^2 = 4x^3 + ax + b.$ Figure 2 shows two examples.

Figure 2: Graphs of Elliptic Curves



They look a little different from the conic curves, but you analyze them in the usual ways.

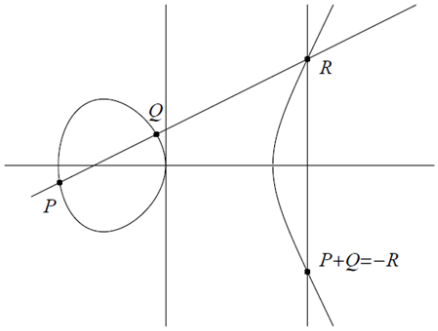
- They are symmetric about the x axis, since every non-zero value for y is accompanied by its negative, as the equation only involves y^2.
- We only get values of x in the case that $4x^3$ + ax + b is ≥ 0, since y^2 is never negative.
- The equation always has one at least one real root, since $4x^3$ + ax + b is positive for large x and negative for very negative x.
- Dividing out that real root, gives a quadratic. If that quadratic has real roots, you get the disconnected graph of Figure 2a, otherwise you get the single root of Figure 2b.

Just as with quadratic equations there is a handy expression called the Discriminant that tells us whether the roots on the right hand side are distinct or not. For the quadratic equation ax^2 + bx + c = 0 the Discriminant D = $b^2 - 4ac$. For our cubic y^2 = $4x^3$ + ax + b the Discriminant D = $4a^3 + 27b^3$.
In both cases the Discriminant is equal to the square of the product of the differences of the roots. So it is non-zero only in the case that all the roots are different.

Thus far this is business as usual. A first indication that there is something more afoot is that the points of the elliptic curve actually form a group. This is stranger than it may appear at first, but if you think about it—there was nothing that said you could take two points of a parabola and put them together to make a third point. Figure 3 (from [3]) shows how the group law works.

Figure 3: Elliptic Curve Group Law

Group Law Geometry



Group Law Equations

$$\lambda = \begin{cases} \frac{y_Q - y_P}{x_Q - x_P} & x_P \neq x_Q, \\ \frac{12x_P^2 - g_2}{2y_P} & x_P = x_Q, \end{cases}$$

$$\mu = \begin{cases} \frac{x_Q y_P - x_P y_Q}{x_Q - x_P} & x_P \neq x_Q, \\ \frac{-4x_P^3 - g_2 x_P - 2g_3}{2y_P} & x_P = x_Q. \end{cases}$$

$$x_{P+Q} = \lambda^2/4 - x_P - x_Q,$$
$$y_{P+Q} = -\lambda r(x_P, y_P, x_Q, y_Q) - \mu.$$

For any two points P and Q on a line that intersects the elliptic curve in three points, the resulting point R of the group operation is the third point on the line—flipped over the X-axis.

Figure 3 also gives the formulas for the group law in terms of the coordinates of the points P and Q. Obviously the formulas are complicated, but the details for now are unimportant. What is important is that there are no radicals in these formulas. If we are thinking about rational points (as with Pythagorean Triples) or even points whose coordinates are in extensions of the rationals (as we saw with Galois theory in chapter 3)—the resulting point R has coordinates in the same field as the coordinates of P and Q. So the group notion for points of an elliptic curve makes sense even if we restrict ourselves to points whose coordinates are rational numbers or numbers in field extensions of the rationals.

Furthermore, if we assume that the coefficients a and b of the elliptic curve $y^2 = 4x^3 + ax + b$ are rational numbers but we look at points with coordinates in a field extension (as in chapter 3), then the elements of the Galois group of the field extension actually permute the points on the curve. That is because when you apply an element of the Galois group to the equation evaluated at a first point, then you get an equation for the coordinates of a second point with the same rational number values for the coefficients. In the same way, the action of the Galois group on points of the curve is actually consistent with the curve's group law in that for any element g of the Galois group, g(P + Q) = g(P) + g(Q). This action of the Galois group on points of the curve turns out to be important for the Modularity Theorem, as we'll discuss later. (It's worth noting in passing that the group law also works for elliptic curves with points in finite fields mod p as in Chapter 1, and those finite elliptic curves are now used in some versions of public key encryption!)

There's one more topic we need to cover before we go on to classification of curves. That also relates to modular arithmetic from chapter 1. One of the ways to deal with solutions of an equation in integers is by examining the solutions to that equation mod p (for prime numbers p). The idea is to try to deduce

information about the original equation by understanding the easier (finite) case of solutions mod p or even mod p^n. It would be nice if the relationship were simple—say that if you could solve mod p for all p then the original equation has solutions. Unfortunately that isn't remotely true—among other examples, the Fermat equation $x^n + y^n = z^n$ for any n has solutions mod all sufficiently large primes! That being said, the technique is still useful and in fact central to the Modularity Theorem, so we summarize what you need to know here.

First of all, by clearing denominators, we can assume all the coefficients of the equation are integers, so that we know what it means to look at the equation mod p. Second, the analysis mod p is cleaner in the case that all the roots of the equation in x are different, so there is a distinction made between "good" primes and "bad" primes. In fact, since the Discriminant tells you whether the roots are different, the "bad" primes are the ones dividing the Discriminant, and everything else is "good". There are various cases for bad primes, discussed under the general topic of "singularities"—we'll show later why that is important but we won't go into detail here. Finally, just to repeat the obvious, the "bad" primes are a finite set.

(As an aside, looking for solutions mod p means typically a change of variables to minimize the Discriminant and hence the number of "bad" primes. In that case the format of the equation may turn out more complicated. The format of the minimal-discriminant equation can be as complicated as $y^2 + a_1$xy + a_3y = $x^3 + a_2x^2 + a_4$x + a_6. With that format, the Discriminant is given by a more complicated formula using the 5 variables a_i.)

Focusing on the "good" primes, the most important result mod p is the total number of solutions. If we let N_p = number of solutions mod p, then standard terminology uses the following equation

$$e_p = p + 1 - N_p$$

The idea here is that on average we expect N_p to be about p + 1 (as we'll see in a minute), so the "error" term e_p measures the departure from the nominal value.

The sense in which p + 1 is the nominal value is given by the following theorem. Remarkably enough there a limit on the size of the "error" that applies for any "good" prime p on any elliptic curve:

$$|p + 1 - N_p| \le 2\sqrt{p}$$ (Haase, 1933)

Qualitatively the value p + 1 also seems a reasonable target, because we can expect roughly half the non-zero values of $4x^3$ + ax + b to be non-squares and half squares. The non-squares will contribute no solutions and the squares 2 solutions (for y and – y). So roughly p/2 values of x will contribute 2 solutions each. Then with some allowance for 0, we end up with p + 1.

In what follows, the error term e_p turns out to be important in a surprising way. Just to reiterate, knowing e_p is equivalent to knowing the exact number of solutions for the elliptic curve mod p.

Classifying Elliptic Curves (Geometric)

Elliptic curves have been studied extensively from the 19th century, and by the beginning of the 20th century there was a quite comprehensive picture of how to classify elliptic curves as geometric objects.

We begin with this classical theory both because it is elegant and interesting in its own right and because of its parallels with the newer classification of elliptic curves as algebraic objects in the Modularity Theorem.[‡]

The difference between the geometric and algebraic points of view came out in the first part of this chapter. In order to find the rational solutions to $x^2 + y^2 = z^2$ we began by looking at an ordinary circle and then found a way to pick out the discrete rational points on the circle. The original circle is a geometric object—we can draw pictures of it with a continuous line representing all of its points. The rational points are something else—they are discrete points picked out from the continuous graph.

In the discussion of Pythagorean triples, the analysis of rational points was something grafted after-the-fact onto the geometric notion of a circle. For Fermat's Last Theorem, as we'll see, the algebraic treatment of elliptic curves is its own subject.

The classical theory deals with complex-number solutions to the degree 3 equations of elliptic curves (see Appendix 1 of the last chapter for background on complex numbers). This is a more general case than for real-numbers, since the real solutions are a subset with imaginary part = 0. It also turns out to easier to work with--since every complex number has a square root, our equation $y^2 = 4x^3 + ax + b$ has y values for every x.

In the classical theory, elliptic curves are associated with parallelograms in the complex plane. Any such parallelogram determines a so-called Weierstrass P function, which maps points in the parallelogram to

[‡] Fermat's Last Theorem is a statement about solutions in integers, and its proof came out of "algebraic geometry," a relatively new area of mathematics which reworks classical geometric results to see what analogous results can be obtained for rational points and other discrete problems. For that reason the relationship between "geometric" and "algebraic" in the remainder of this chapter is different than what we saw earlier. The algebraic results are not grafted on; instead the classical theory of elliptic curves is presented as a model, which then gets reworked to deliver the algebraic results.

It is hard to resist a little parenthetical philosophy on the importance of discrete problems. Although number theory has been around forever, the emphasis on discrete problems is really in the air. Classical science was concerned largely with continuous processes: Newtonian mechanics, Maxwell's equations for electromagnetism, etc. In contemporary science—quantum mechanics, computer science, etc.—the point of view is predominantly discrete. That applies to electronics, nuclear physics, and the astronomy of black holes and neutron stars. Discrete problems are hard, and classifying elliptic curves, as a step beyond conics, represents a fundamental advance in understanding.

It should also be noted that classification theorems of this sort are extremely important in mathematics. In any kind of application, you only know what is possible when you know what can mathematically exist. Another recent example of a monumental classification theorem is the 1955 – 2004 classification of finite simple groups—which enumerated what kinds of finite groups can exist. This is another apparently strange result: 3 well-known infinite classes of groups and then exactly 26 exceptional cases, including one group—called the "monster"--with more than 10^{53} elements! This is not just of theoretical interest, as these groups represent important symmetries in physics.

an elliptic curve. Specifically, the P function and its associated derivative P' (another function you can get from P itself) give you every point on the curve by mapping the values of a complex number t to the point $(P(t), P'(t))$, i.e. x = $P(t)$ and y = $P'(t)$. What is more, the mapping actually sends sums of t-values to addition of points via the group operation on the elliptic curve, that is

$$(P(t_1 + t_2), P'(t_1 + t_2)) = (P(t_1), P'(t_1)) + (P(t_1), P'(t_1)),$$

where addition on the right-hand side is the group law we saw defined for the elliptic curve.

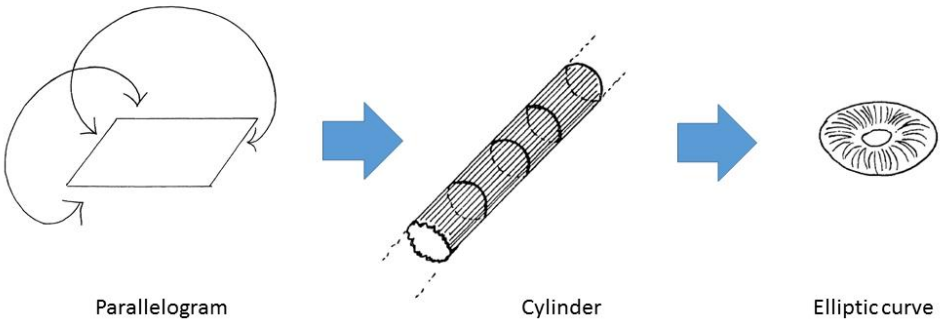
Finally there is a standard function "j" of the coefficients a and b of the elliptic curve $y^2 = 4x^3 + ax + b$ that tells you exactly which curves are considered equivalent (i.e. related by change of coordinates). The value is given by $j = \frac{1728\,a^3}{D}$, where D is the Discriminant we discussed earlier. All curves you get by inflating the parallelogram (multiplying both complex numbers by the same value) or more generally by rational changes of coordinates[§] are equivalent and have the same value for the j function.

In the classical theory the parallelogram has a nice geometrical interpretation, but you have to think about it a bit because it's all based on complex numbers. If we look at solutions to $y^2 = 4x^3 + ax + b$ where x and y are complex numbers, then the graph is a 2-dimensional surface instead of a line (since the complex numbers themselves are represented as a plane and not a line). That surface (with a mathematically-added point at infinity) turns out to be a torus, the surface of a doughnut. And you can think of the torus as a parallelogram with pairs of opposite sides are identified with each other (see Figure 4—from [5]).

(That "point at infinity" is a little hard to get your arms around—and we won't belabor it here--but we can at least reconcile the two graphs in Figure 2 with the torus in Figure 4. The first thing to remember is that each graph in Figure 2 shows real number points—i.e. a line—whereas the torus represents complex number points—a surface. The real-number points on the surface are cut out by a plane passing through the surface. If you imagine a point on the right side of a torus as exploded out to real infinity, then the two graphs in Figure 2 are cut out by planes passing through the torus horizontally. In Figure 2A the plane goes through the donut hole, in Figure 2B it goes through the torus but passes above or below the hole. The point at infinity itself comes from looking at the curve in projective geometry; for detail see the Appendix to [11].)

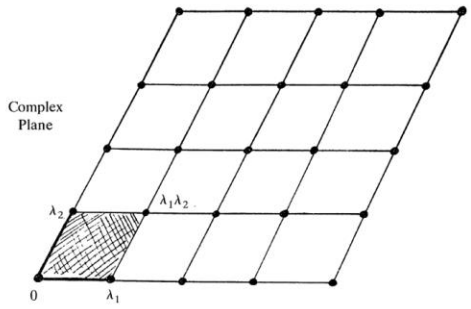
[§] Specifically a "rational changes of coordinates" means that the change of coordinate functions are quotients of polynomials. For curves to be equivalent you need to have change of coordinates functions in both directions that undo each other when you do them back-to-back.

Figure 4: Parallelogram and Torus



Finally, if you put the parallelograms together, you can fill out the whole complex plane with the parallelograms as tiles (see Figure 5—also from [5]). Any function that acts the same on each tile is considered to be a function on the elliptic curve. Such a function is "doubly periodic", because its value repeats as one moves in the direction of either side of the parallelogram. The Weierstrass P-function is in fact defined everywhere and is doubly-periodic in this way.

Figure 5: Tiling by Parallelograms



So in the end the geometric classification of elliptic curves boils down to something rather straightforward:

A pair of complex numbers (that don't lie on the same line to the origin) determine a parallelogram in the complex plane. They also determine a Weierstrass P function that maps the points of the

parallelogram to the points of an elliptic curve satisfying an equation of the form $y^2 = 4x^3 + ax + b$. Functions on the curve are defined to be functions in the complex plane that are invariant under translations by the sides of the parallelogram. All elliptic curves can be obtained in this way, and two curves are equivalent if and only if they have the same value for the j function.

Before leaving this section we should say a little more about the rather magical j function—which turns up in very many contexts (including this example involving both theoretical physics and the "monster" group mentioned earlier). First of all, because inflating a parallelogram gives equivalent elliptic curves, the j function depends just on the ratio of the sides of the parallelogram, so it can be viewed as a function of a single variable z (i.e. the ratio) on the upper half plane of the complex numbers. Furthermore, with this variable z, parallelograms turn out to be equivalent if and only if they are related by changes of variable of the form

$$z \to \frac{az+b}{cz+d}$$

where a, b, c, and d are integers with ad – bc = 1. That means the j function is independent of such changes of variable, that is

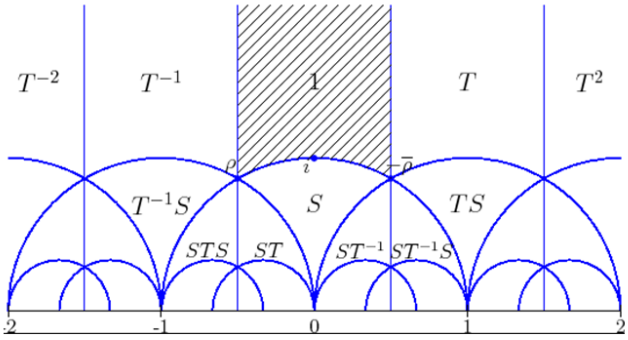
$$j\left(\frac{az+b}{cz+d}\right) = j(z).$$

Changes of variable of this sort form a group, called the "modular group." Elements of the modular group are represented as 2 x2 matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with matrix multiplication as the group operation. Functions that are independent of the action of the modular group—as we just saw for the j function—are called "modular functions." (For the record there are a few other technical restrictions on which functions can be modular, but they're not relevant here.)

Since the modular group acts on the upper half plane, we can talk about tilings for the modular group. That is, we can look for geometric tiles filing the upper half plane, where each tile represents a full set of distinct points for the modular group. Figure 6 shows the result. Any tile can be transformed to any other by an element of the modular group, and the tiles as a whole cover the upper half plane exactly once.

Figure 6 is interesting, because it is visual representation of the classification of elliptic curves. Each z value represents a ratio of parallelogram sides and therefore a set of elliptic curves. The z values correspond to equivalent curves if and only if they are related by the action of the modular group. So each tile is a complete set of equivalence classes of elliptic curves. Furthermore, since j values are the same if and only if the curves are equivalent, the j function takes distinct values on each point of the tile. And in fact every complex number arises as a j value.

Figure 6: Tiling for the Modular Group



In the figure each of the vertical strips and the smaller curved triangular sections is a tile for the modular group. The tiles can look different from each other, since the modular group includes elements that are not just ordinary translations. The figure makes this transformation of tiles explicit, because the "S" and "T" labels on particular tiles refer to particular elements of the modular group that were applied to the shaded region to get there. In the modular group the element S is the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ corresponding to $z \rightarrow -1/z$ and element T is the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ corresponding to the translation $z \rightarrow z+1$. Those two elements, multiplied out, generate the entire modular group.

As you can see by now, the classical theory gives a very precise description of the possible elliptic curves. However since its whole underpinning is geometric (the P function itself involves Calculus), it doesn't directly yield results about classifying solutions in integers, rational numbers, or other discrete fields. To address that additional level of detail it was necessary to develop a new algebraic classification of elliptic curves using methods that no longer rely on continuous geometry. The classical case is a good introduction to the algebraic case, because there are many parallels (including a second coming of the modular group). However, as we'll see, the new classification is an intriguing combination of ideas old and new.

Classifying Elliptic Curves (Algebraic)

The algebraic classification deals specifically with rational elliptic curves, i.e. curves that are defined by equations with rational numbers as coefficients. Many of the key concepts grow out of what we just talked about for the classical case, but as you'll see they get applied in very different ways.

To start with, we need a couple of extensions to the concepts we just defined for the modular group:

1. In addition to modular functions, we need to define one other type of function related to the modular group—a so-called "modular form". Instead of being independent of the modular group, a modular form varies in a specific way:

$$f(\frac{az+b}{cz+d}) = (cz+d)^2 f(z) .$$

This seems rather arbitrary as a definition, but in fact modular forms of this type are used to create objects called "differentials" (a version of the differentials from Calculus), and this type of dependence on the modular group makes the differential as a whole independent of the modular group. Differentials themselves are beyond the scope here, but modular forms turn out to be crucial, as we'll see in a minute.

2. The modular group has important subgroups. Most of the action is with so-called "congruence subgroups" where the value "c" in the matrix is = 0 mod some integer N. (Notice that in the product $\begin{pmatrix} a1 & b1 \\ c1 & d1 \end{pmatrix}$ x $\begin{pmatrix} a2 & b2 \\ c2 & d2 \end{pmatrix}$ the new "c" term is c1 x a2 + d1 x c2 = 0 mod N, so the product of two such matrices really is in the subgroup.) Standard notation for the subgroup of the modular group with c = 0 mod N is Γ_0(N). Since we now have multiple groups to consider, modular functions or modular forms will now always be defined with respect to a specific group Γ_0(N) under consideration.

With that we have what we need to begin describing the modular classification of elliptic curves. As in the classical discussion we will define tiles and mappings to elliptic curves. However, remarkably enough, the parallelograms from the classical theory get replaced by something closely akin to the j-function tiles from Figure 6!

1. The congruence subgroups Γ_0(N) (i.e. the two-by-two matrices of integers $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where ad – bc = 1 and c = 0 mod N) form the basis for the classification, because they provide the first step in constructing models for elliptic curves. For each such congruence subgroup there is a corresponding constructed curve—called X_0(N)—that fits into a picture much like Figure 4. That is, there is a tiling of the upper half-plane for the group Γ_0(N) so that any single tile gets mapped to the curve X_0(N).

As noted earlier, however, the details of this picture are a little different from Figure 4. In that previous picture we had parallelogram tiles mapped to a torus representing an elliptic curve. Tiles were related to each other by translations in two directions. In our current case the tiles instead look much like the ones in Figure 6. The reason is that although the matrix S = $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ from Figure 6 is not in Γ_0(N) for N > 1, a similar matrix $\begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$ is—which is why the tiling turns out to be a rescaled version of Figure 6. So we get tiles that look like the ones in Figure 6 mapped to this new constructed curve X_0(N). We get one such mapping for every value of N. And in this case the mapping is purely algebraic—no Calculus as there was with the P function.

2. However it turns out that this time we're only half-way there. The reason is that the constructed curve is not always an elliptic curve. (Topologically it will look like a "sphere with handles", where the torus is the case of a single handle. Most of the time there will be more than one handle.) We need another step to get all the way to models for elliptic curves.

To finish we need to look also at modular forms for our congruence subgroup Γ_0(N). Recall these are functions that satisfy

f($\frac{az+b}{cz+d}$) = $(cz+d)^2$ f(z) for every matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ in Γ_0(N)

Recall also that such modular forms correspond to "differentials" defined on the constructed curve X_0(N). It turns out that the number of such differentials is equal to the number of "handles on the sphere" for X_0(N), and they can be used to construct elliptic curves from X_0(N) itself. Specifically, certain modular forms f for the congruence subgroup Γ_0(N) give a mapping from X_0(N) to a constructed elliptic curve—usually called A_f. We get one of these for each combination of a number N and an appropriate modular form f. Later we'll say more about which modular forms can be used and how their properties relate to the properties of the constructed elliptic curves. Again the constructions are purely algebraic.

To summarize:

1. The classification of elliptic curves is based on the congruence subgroups Γ_0(N).

2. Each congruence subgroup gives us two types of items:

   - a specifically-constructed curve X_0(N) that is mapped from tiles for Γ_0(N). We get one of these for each number N.

   - one or more modular forms f defined for the congruence subgroup Γ_0(N).

3. Certain of these modular forms f—together with the curve X_0(N)—are used to construct elliptic curves A_f. These A_f curves are the models that turn out to represent all elliptic curves.

As context, it is useful to understand what is old and new here. Modular functions and forms have a long history, as families of those functions have been much studied from the 19th century for their value in proving theorems in number theory. There are formulas to count them in terms of the properties of the curves X_0(N) as geometric objects. However the constructions here of X_0(N) and A_f rely on algebra rather than Calculus. Most of the properties we will discuss from here on only makes sense in terms of the algebraic approach.

To be clear, we have thus far described a method to generate a specific set of elliptic curves by an algebraic construction from congruence subgroups of the modular group. That construction preceded the Modularity Theorem by decades, and as we will see the curves created in that way have some remarkable properties. What the Modularity Theorem established was that those constructed curves are in fact universal. Up to change of coordinates they are models of everything.

In what follows we will first focus on the construction—the modular forms f and the properties of the constructed elliptic curves. Then we will move on to the Modularity Theorem proper—what it means for the constructed elliptic curves to be universal.

While the logic that constructs the elliptic curves A_f is beyond our scope, there is still quite a lot to say about their properties. In particular we will be specific about which modular forms perform this role and how they are related to the constructed curves.

The first important property of the curve A_f is that the primes dividing the Discriminant of A_f are precisely the ones dividing N. So the "good" primes for A_f are precisely the ones that don't divide N. However, that is just the start of what happens with the "good" primes of A_f.

The crux of the connection between the modular form f and the associated elliptic curve A_f involves the values e_p that we saw in our discussion of elliptic curves mod p. In that previous discussion the value e_p was defined to (p + 1) – (number of solutions mod p), i.e. the difference between the exact number of solutions and the assumed average value p + 1. We will be looking at these values for the "good" primes of A_f

To understand the relation of the modular form to the curve we first have to talk a little more about modular functions and forms in general. First of all, every group Γ_0(N) contains the matrix $\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$ corresponding to the translation z → z+1, because the c value is actually 0. What's more, for this matrix the modular form equation

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^2 f(z)$$

reduces to simply

$$f(z+1) = f(z)$$

because c = 0 and a, b, and d = 1. In other words, every modular function or form is a periodic function, like a sine or cosine but with period = 1.

It is a theorem in Calculus that every periodic function with period 1 can be expressed as a possibly infinite sum of terms $\sin(2\pi nx)$ and $\cos(2\pi nx)$ with appropriate coefficients. This is a Fourier series, as discussed in Appendix 3 of chapter 1. Here however we can simplify the format of the series, since we know from Euler's formula that e^{ix} = cos x + i sin x. The result is that we can write the Fourier series in the following simple form with q = $e^{2\pi ix}$ and complex coefficients a_n:

$$f(z) = \sum_{n=1}^{\infty} a_n\, q^n$$

Since any multiple of f will have the same behavior, we will assume this form is normalized by setting a_1 = 1. (The Fourier series could also have had an a_0 term also, but in fact the classification is only concerned with so-called "cusp" forms where a_0 = 0. There are also other technical restrictions on f that are beyond the scope here.)

With those preliminaries done, we can now say which modular forms will be involved in the classification—for the classification all of the coefficients $a_n$must be integers. That is to say for each modular form f with integer values for the Fourier coefficients a_n we get an elliptic curve A_f in our classification.

With this association of modular forms and elliptic curves A_f we get an astoundingly intimate relationship between the curve and the Fourier series $\sum a_n q^n$ of the modular form f:

$a_p = e_p$ (as defined for the curve A_f) for every prime p not dividing N

Think about it--the modular form has embedded in it the number of solutions of A_f mod p for every "good" prime of the elliptic curve A_f! (Note that since the e_p values come from counting solutions, it makes sense that we restricted attention to modular forms f where the a_n values are integers.)

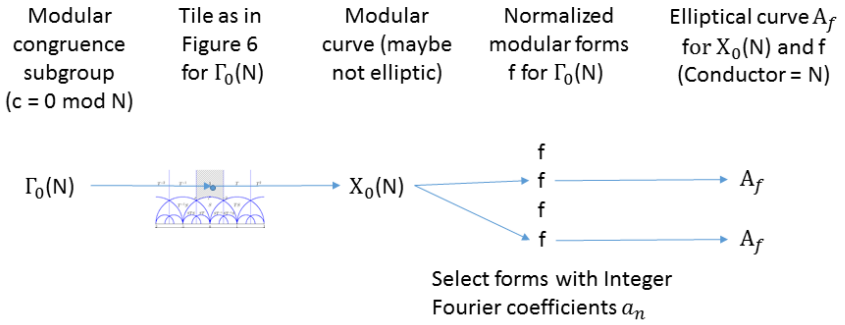
The explanation for why this happens goes far beyond the scope of this chapter. What is of interest, however, is how closely this is tied in with the notion of a Galois group from chapter 3.

Earlier in this chapter we saw how a Galois group can permute points of an elliptic curve with coordinates in field extensions of the rational numbers. That ties into our current discussion in a very specific way. For any number n, you define the "n-division points" of an elliptic curve to be the points P where P + P + … + P (P added n times using the elliptic curve group law) = the identity for the group law. Since we're adding a point n times, the formulas for the group law (iterated n times) give an equation satisfied by the coordinates of any n-division point. Solutions of that equation will include new numbers extending the rational numbers, just as in Chapter 3. And the elements of the Galois group for the field extension permute the division points.

This simple idea of division points permuted by a Galois group has enormous impact. On one hand, it turns out that these Galois group permutations of n-division points (for varying n) determine the number of solutions for the elliptic curve A_f mod p (for good primes p)—and hence the values e_p. And on the other hand, the actions of Galois group elements turn out to correspond to well-studied operations on modular forms—and hence the values of a_p. So the notion of Galois group is the glue that brings together the two sides of $a_p = e_p$.

We now give a picture of the classification to this point in Figure 7.

Figure 7: Modular Correspondence (1)



There is one new feature in Figure 7 that deserves mention. We previously noted that the Discriminant of A_f has the same primes as N. There is in fact another, more sophisticated object called the "Conductor" we can associate with any elliptic curve. The Conductor has the same primes as the minimal Discriminant, but the exponent for each prime p depends on the form of the curve mod p (this is the "singularity" class discussed earlier). In the case of the elliptic curve A_f the Conductor exactly matches the value N.

The Modularity Theorem

The Modularity Theorem finishes the job of the modular classification by showing that every rational elliptic curve is equivalent to a constructed elliptic curve A_f. Figure 8 adds this last stage to the picture in Figure 7.

There is an important new term "isogeny" in Figure 8. This defines what it means for the elliptic curves A_f to be universal by saying what it means for two curves to be algebraically equivalent. The definition is as follows.

Two elliptic curves are "isogenous" if there is a non-constant algebraic map of curves from one to the other that takes the identity element in one to the identity element of the other. It turns out that such a map necessarily preserves the group operation, and further that there will actually be such a map in each direction from one curve to the other. This can be viewed as a generalization of a change of coordinates—it means that the algebraic properties of the two curves are the same.

Isogenous curves turn out to have the same Conductor and also the same e_p values for primes outside the Conductor. What is more (a non-trivial result of Gerd Faltings) the converse is also true—two

elliptic curves with the same e_p values outside the Conductor are in fact isogenous. This is an interesting example case where properties mod almost all primes imply a general result (isogeny) for the curves themselves.

Figure 8: Modular Correspondence (2) – Modular Theorem

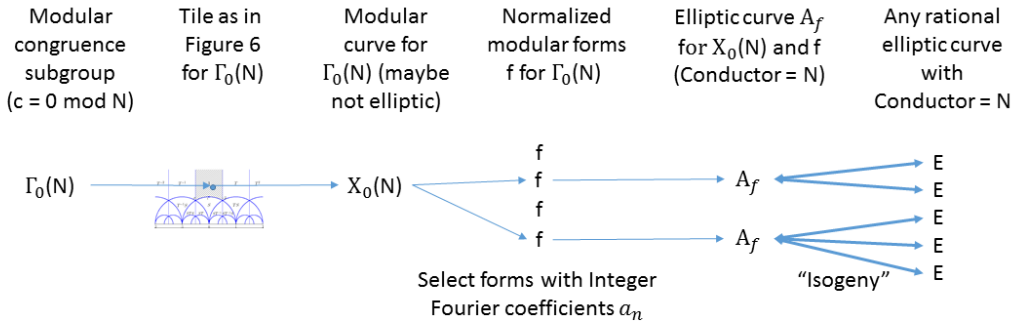


Figure 8 is remarkable from many points of view:

- The curves X_0(N) and elliptic curves A_f are explicit constructions that only after the fact turned out to be universal.

- We can now say that for every elliptic curve with rational coefficients there is a modular form somewhere that knows the number of solutions of the curve mod p for every p outside the Discriminant.

- By construction the modular classification provides considerable data about each curve, so that it is a powerful tool in analyzing elliptic curves and other algebraic questions. In that sense the proof of Fermat's Last Theorem, as an unexpected spin-off, is an indication of how useful the classification can be.

While the proof is far outside the scope of this chapter, it is interesting to use its results to see what it means for a given elliptic curve to be modular. (One thing we can say about the proof is that the Galois group figures here again. By Falting's theorem it suffices to find a modular form matching the e_p values, and in practice that means matching the Galois action on "division points" to corresponding operations on modular forms.)

A given elliptic curve fits into the classification as follows:

1. From the curve itself you calculate the Conductor for the minimal-discriminant equation. That gives the value N.

2. There is a modular form for Γ_0(N) whose Fourier coefficients match the e_p values for every "good" prime of the curve. (This modular form is a so-called newform, which means it is not a modular form for any Γ_0(M) where M divides N. By definition Γ_0(M) $\supset$ Γ_0(N) in this case, so any modular form for Γ_0(M) is automatically a modular form for Γ_0(N). The matching form is "new" in the sense that it doesn't come from a smaller value of N.)

3. That modular form can be used to construct an elliptic curve A_f. The given elliptic curve is isogenous to A_f.

4. Through the isogeny there are modular functions f and g for the group Γ_0(N) such that x = f(t), y = g(t) give all the points of the curve--as with the Weierstrass P and P' functions.

To make matters more concrete, we give an example taken from reference [7]. The simple equation

y^2 + y = x^3 - x^2 has Discriminant divisible only by 11. And the modular form

$$X\prod_{m=1}^{\infty}(1-X^m)^2(1-X^{11m})^2=\sum_{n=1}^{\infty}a_nX^n$$

has a Fourier expansion where a_p = e_p for every p ≠11.

Fermat's Last Theorem

We're now ready to apply the Modularity Theorem to Fermat's Last Theorem. For this we assume there is a counterexample A^p + B^p = C^p. Since we can remove any common factors of A, B, and C, we assume that has been done.

Using these values we consider the surprisingly simple elliptic curve y^2 = x(x − A^p)(x + B^p). Since the Discriminant is the square of the differences of the roots of the right-hand equation (i.e. 0, A^p, $-B^p$) we have

$$\mathrm{D} = A^{2p}B^{2p}(A^p + B^p)^2 = A^{2p}B^{2p}C^{2p}$$

This is a strange-looking pure 2pth power, but the contradiction we're looking for is not obvious.

To apply the Modularity Theorem we first need to compute the Conductor N of this curve. For this case, that turns out to be the product of all the primes in D, each taken to the first power. Notice that one of A, B, C has to be even, so 2 (but not 4) is a factor of the Conductor. Overall the Conductor has single powers of primes but the Discriminant has every prime to the 2p power or a multiple.

Now by the Modularity Theorem we can find a modular form f for Γ_0(N) whose Fourier coefficients match the e_p values for the curve. Further f is a newform so it doesn't come from any M dividing N. However it turns out (very non-trivially) that in this case where the pth power of each odd prime divides the Discriminant—but only the first power divides the Conductor—we can divide out all the odd primes from the Conductor and find a (different) modular form in Γ_0(N/(product of odd primes)) = Γ_0(2).

But $\Gamma_0(2)$ is a specific, well-studied group, and there are no appropriate modular forms for it. (Notice that for $\Gamma_0(N)$, the smaller the value of N, the bigger the group, and therefore the fewer modular forms that satisfy the constraint.) And that contradiction proves Fermat's Last Theorem!

Names

Without diminishing the achievement of Andrew Wiles, using any single person's name understates the magnitude of the work represented by the Modularity Theorem and its application to Fermat's Last Theorem. For that reason we end this chapter with an annotated version of the previous historical summary.

Historically, the Modularity Theorem goes back to conjectures first formulated in the 1950's, which became mainstream by 1960's (Yutaka Taniyama (Japan), Goro Shimura (Japan), Andre Weil (France)). By the 1970's it was one of the major activities in mathematics, involving a highly-international cast of characters (the bibliography in the technical overview [7] has 61 names from 11 countries). The connection to Fermat's Last Theorem was first conjectured in 1985 (Gerhard Frey (Germany) using a curve first mentioned by Yves Hellegouarch (France)) and then established in 1986 (Jean-Pierre Serre (France), Ken Ribet (US)). The proof of the Modularity Theorem itself (in the cases needed for Fermat's last theorem) was announced by Andrew Wiles (UK) in 1993, but a gap was found in the argument. After more than a year of frantic effort the gap was finally closed by Wiles assisted by his former student Richard Taylor (UK)—and that proved Fermat's Last Theorem. Wiles' methods were then extended to cover all cases for the full Modularity Theorem in 2001 (Christophe Breuil (France), Brian Conrad (US), Fred Diamond (US), Richard Taylor (UK)).

Since Taniyama got this ball rolling, Shimura's tribute to his early-deceased friend [9] seems relevant. And as a final note, in any effort of this size assessing a single person's contribution is complicated—and can be contentious. In that light, anyone who thinks mathematics is a dispassionate endeavor should read [4]!

References for this Chapter

This list of references shows where to get more detail—at various levels of difficulty. The most accessible are [2], [5], and [8]. [1], [10], and [11] are approachable treatments of related but more general subject matter. [7] was particularly helpful to writing this chapter. [3] and [6] are for the brave.

[1] Avner Ash and Robert Gross, Fearless Symmetry, Princeton University Press, 2006

[2] David Cox, "Introduction to Fermat's last theorem," American Mathematical Monthly, **101** (1), January 1994, pp. 3–14

[3] Fred Diamond and Jerry Sherman, A First Course in Modular Forms, Springer 2005

[4] Serge Lang, "Some History of the Shimura-Taniyama Conjecture," Notices of the AMS November 1995, http://www.ams.org/notices/199511/forum.pdf

[5] Barry Mazur, "Number Theory as Gadfly," American Mathematical Monthly August-September, 1991, https://www.maa.org/sites/default/files/pdf/upload_library/22/Chauvenet/Mazur.pdf

[6] J. S. Milne, Elliptic Curves, Kea Books, 2006

[7] Kenneth Ribet, "Galois Representations and Modular Forms," Bulletin of the American Mathematical Society, October 1995

[8] Kenneth Ribet and Brian Hayes, "Fermat's Last Theorem and Modern Arithmetic," American Scientist, March-April 1994

[9] Goro Shimura, "Yutaka Taniyama and his time," Bulletin of the London Mathematical Society 1989, http://blms.oxfordjournals.org/content/21/2/186.full.pdf

[10] Joseph Silverman, A Friendly Introduction to Number Theory, Pearson, 2012

[11] Joseph Silverman and John Tate, Rational Points on Elliptic Curves, Springer 2015