An Introduction to Mathematics Beyond Calculus

Mathematics 17 — Winter 2012

February 12, 2012

Thomas R. Shemanske

Contents

Preface. The typical experience of a potential mathematics major who survives the calculus sequence with enough curiosity remaining to continue their study of mathematics, is to explore discretely the broad array of topics which constitute modern mathematics: linear and abstract algebra, real, complex, and functional analysis, algebraic and differential topology, geometry, probability, statistics, combinatorics, logic, set theory, number theory, as well as a myriad of focused applied subjects.

Unfortunately, to the undergraduate this often appears to be exactly what was intimated above: discretely chosen, if disconnected, subjects. However this is far from the truth. As one takes more advanced undergraduate and certainly graduate courses, one begins to see the interconnections between many of these subjects which makes mathematics so vibrant a subject.

It is the goal of this course to attempt to present some of the vista of modern mathematics, and some of the interconnections between some of these subjects at an early stage in your careers, which hopefully will invigorate your view of mathematics and its potential.

The broad theme of this offering of Math 17 is the ubiquity of algebraic structures in mathematics and perhaps even in the world at large. We will develop a number of basic and pervasive algebraic and number theoretic tools with an eye to intertwining applications in algebra, geometry, cryptography, and number theory. A focal point will be to understand how certain curves in the plane (elliptic curves) have both algebraic and geometric connections to problems as diverse as Fermat's Last Theorem to major efforts in modern cryptography.

CHAPTER 1

Three Motivating Problems

We consider three problems in number theory and geometry whose solutions use elliptic curves in an essential if sometimes subtle manner, and which we use to motivate the study of elliptic curves. Two of the problems, Fermat's Last theorem and the Congruent Number problem, are problems whose statements the ancient Greek's would have understood. In contrast the third, elliptic curve cryptography, is quite modern, and about which Koblitz et al. [**3**] write that over the last sixteen years, "elliptic curve cryptography went from being an approach that many people mistrusted or misunderstood to being a public key technology that enjoys almost unquestioned acceptance."

What are these elliptic curves which have had such an impact and why?

At first blush elliptic curves appear to be nothing special. In secondary school it is common to study plane and analytic geometry, to learn about properties of lines and conics in the plane as well as the principles of Euclidean geometry. As we start our study, we consider a cubic curve, one which could easily have been studied in secondary school.

A cubic curve has the form

$$ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 + hx + iy + j = 0,$$

and an elliptic curve is a special case of a cubic curve, generally in the form

$$y^2 = x^3 + ax^2 + bx + c,$$

where the cubic $x^3 + ax^2 + bx + c$ is non-singular (has distinct roots).

While the definition of an elliptic curve seems to shed no light on their importance, their impact on the solutions to major questions in mathematics as well as providing some of the best schemes for public key cryptography suggests a closer look.

We shall look only briefly at the Fermat problem, only slightly more deeply at the Congruent Number problem, and most deeply at developing an understanding of how elliptic curves are of critical use in cryptography. This emphasis is deliberate, in part because the role elliptic curves play in the solutions of the Fermat and Congruent Number problem is more subtle and sophisticated, and in part because such an omission provides the opportunity for students to consider these problems more deeply for their term projects.

We shall spend the majority of the term trying to come to grips with the basic properties of elliptic curves and their applications to cryptography.

1. Fermat's Last Theorem

This theorem, conjectured by Fermat in 1635, states simply that for $n > 2$ the equation $x^n + y^n = z^n$ has no solutions in the integers except when one of the variables is zero. We note that this contrasts sharply to the case of $n = 2$ for which solutions (Pythagorean triples) abound. In fact, Pythagorean triples will play an integral role in the Congruent Number problem, but before jumping ship, we say a few more words about the Fermat theorem.

In 1640, Fermat himself proved that the conjecture was true for $n = 4$, and noted that if $n = km$, a non-trivial solution to $x^n + y^n = z^n$, means nontrivial solutions to $(x^k)^m + (y^k)^m = (z^k)^m$, that is to a Fermat problem whose exponent is a divisor of the original. Fermat's $n = 4$ result and this observation reduces the proof of the Fermat theorem to showing there are no nontrivial solutions when $n = p$ is an odd prime.

Until 1839 only the cases $n = 3, 5, 7$ had been resolved; this included Sophie Germain's important work which eventually allowed the conjecture to be proved for all odd primes less than 100. In the 1850's, Ernst Kummer developed techniques to prove the Fermat conjecture for all "regular" primes, which is believed to be an infinite family. Modern computing methods verified the theorem for primes less that four million.

The first real breakthrough came in 1985 when Gerhard Frey suggested that a counterexample to Fermat's theorem could be used to create an elliptic curve having properties which would provide an counterexample to yet another unproved conjecture due to Taniyama and Shimura. This was very disquieting!

In the period 1985-86, Jean-Pierre Serre showed how the Taniyama-Shimura conjecture together with another smaller conjecture — termed the "epsilon conjecture" — would imply Fermat's theorem. Ken Ribet proved the epsilon conjecture in 1986 reducing the Fermat theorem to a proof of the Taniyama-Shimura conjecture for a special class of elliptic curves.

In 1994 Andrew Wiles (after 7 or more years of intense work, a together with a last minute save by Richard Taylor) succeeded in proving the required case of the Taniyama-Shimura conjecture, and hence proving Fermat's Last theorem.

And in case you were wondering, the Taniyama-Shimura conjecture has now been proven as well.

2. The Congruent Number problem

A positive integer is called a *congruent number* if it is the area of a right triangle whose sides all have rational length. For example, 6 is a congruent number since 6 is the area of a 3-4-5 right triangle.

It is also true that 5 is a congruent number, though this is something you might not guess right off. But indeed, 5 is the area of the right triangle with sides: 3/2, 20/3, 41/6. Any reasonable person would agree that one can check the result, but it certainly is mysterious how one would come up with a triangle having those sides. Indeed it now becomes a much more interesting question to ask which integers are congruent numbers.

A key observation is that if N is a congruent number, then so is Nt^2 for any positive integer t; indeed if N is the area of a triangle having rational sides a, b, c, then Nt^2 is the area of a triangle with sides at, bt, ct. Let's consider the triangle showing that 5 is a congruent number. It is easily seen that 6 is the common denominator of the rational numbers $3/2$, $20/3$, $41/6$, and from our observation above $5 \cdot 6^2$ is also a congruent number, being the area of a right triangle with sides $9, 40, 41$. But of course this means that 9, 40, 41 is a Pythagorean triple! Conversely, suppose that N is the area of a right triangle with integer sides A, B, C. Write $N = N_0 t^2$ where N_0 is square-free. Then N_0 is a congruent number, being the area of a right triangle with rational sides $A/t, B/t, C/t$.

So there is a clear relationship between congruent numbers and Pythagorean triples, which means if we had a way to list all Pythagorean triples, we would know which numbers were congruent numbers. In fact, we will show how to list all the Pythagorean triples. Unfortunately, the congruent numbers that come out of the list are not in order and are often repeated, so this procedure cannot definitely answer whether a given integer, say 157, is a congruent number. Still it will provide a good deal of insight, and we will spend some time with it.

On the other hand, Jerrold Tunnell developed a condition based on the arithmetic of elliptic curves which provides a beautiful answer to this question. Perhaps someone will explore this for a term project.

3. Cryptography

This is a subject having a long and fascinating history, and a matter of crucial importance to all of us in an age when so many transactions happen electronically. It is the subject around which essentially all the background material on number theory and algebra which we develop will be focused.

There are many interesting questions of a practical nature which cryptography solves, and which we shall examine, but as a teaser, we mention a few in this introduction. It is not terribly difficult to send private messages to a friend even over an insecure channel. What becomes trickier is when you want to do the same with someone you don't know. Why would you want to do that? Well, every time you order something online, you want to communicate securely with the vendor so that confidential information (e.g., credit card number) is not revealed over the insecure web. But how can you (read your computer) and your vendor do this? How can someone who has received an email from you prove to a third party that the message is indeed from you and not someone forging your address? How can someone be sure a message has not been tampered with (e.g., please transfer $XXX from Dartmouth's payroll account to my personal account).

All of these are vital questions which modern cryptography answers effectively, and elliptic curves figures prominently in the mix. Of course given any cryptographic system, there are many individuals who do their best to break it, so we will look at standard kinds of cryptographic attacks on various systems, and how vulnerable each system is to different types of attacks. In the end, elliptic curve cryptography turns out to be among the best public key systems currently in use.

CHAPTER 2

Back to the beginning

So far, we have seen some number theory popping up, but one should ask where is all the advertised interplay of algebra, geometry, and number theory? Let's start with some geometry and an attempt to broaden your perspective. Let's start with an innocent looking question: What are the solutions to $x^2 + y^2 = 1$?

Well of course geometrically we think of the solutions to this equation as the coordinates of points in the plane on the unit circle. If you are fresh out of a calculus course, you might even say the points on the circle are parametrized by cosine and sine, that is each point (x, y) on the unit circle has the form $(\cos\theta, \sin\theta)$ for a uniquely determined $\theta \in [0, 2\pi)$.

This is certainly one reasonable answer. Now let's ask what are the rational points, that is what are the points (x, y) on the unit circle both of whose coordinates are rational numbers? Admittedly, this seems like an odd question to ask.

Well, let's see: Suppose that $\left(\frac{a}{b}\right)^2 + \left(\frac{c}{d}\right)^2 = 1$. Then $(ad)^2 + (bc)^2 = (bd)^2$, that is a rational point on the unit circle corresponds to a Pythagorean triple. What an interesting coincidence. Conversely, any Pythagorean triple $A^2 + B^2 = C^2$ with $(A, B, C \in \mathbb{Z},\ C \neq 0)$ corresponds to rational point $(A/C, B/C)$ on the unit circle. And note that the constrain $C \neq 0$ is not much of a constraint, since if $C = 0$, the only possible triple of real numbers is 0, 0, 0 which isn't all that interesting.

So let's recap from the first chapter to this point: congruent numbers are connected to Pythagorean triples, and Pythagorean triples are connected to rational points on the unit circle $x^2 + y^2 = 1$. We knew if we could list all the Pythagorean triples, all the congruent numbers would eventually come out, so now all we need to do is figure out how to characterize all the rational points on the unit circle. Before addressing that question, let's wander a bit more.

What if we generalized the problem of asking not simply for the solutions to $x^2 + y^2 =!$, but for solutions to $x^2 + y^2 = r$? We might be led to consider solutions to $x^2 + y^2 = 0$ or $x^2 + y^2 = -1$. If we were biased by our interpretation above where we sought only solutions in the rational or real numbers, we might be inclined to say the only solution to the first equation is the origin, and that there is no solution for the second. But that is only because our perspective may be a bit narrow. For example if we were looking for the solutions in the complex numbers or over a finite field (whatever that is), there would be lots of solutions. That flexibility will be very useful to us.

1. Geometry versus Algebra

When we think of solutions to equations, our perspective is influenced by the underlying ring or field in which we seek solutions. For example when we consider $x^2 + y^2 = 1$ over the real numbers, we see the unit circle in the plane. If we asked for only the rational points (well we haven't found them all yet, but for sure), there would be far fewer [countably versus uncountably many], and if we looked for solutions in the integers, there would only be four.

When geometric interpretations of solution sets do not prove insightful, we often look at any algebraic structure which the solution set may have. Let's start with something having both a geometric and algebraic structure. Let

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x + y + z = 0\}.$$

Of course from calculus, we recognize this as a plane through the origin with normal vector (parallel to) $\langle 1, 1, 1\rangle$, but the set also has an algebraic structure. The set S is nonempty, since $(0, 0, 0) \in S$, and we note that the sum or difference of two points in S is again in S, as is any scalar multiple of an element of S. This makes S into what is called a vector space, in particular, a subspace of $\mathbb{R}^3$. Note the same is not true for planes of the form $x + y + z = k$ for nonzero k.

Also note that every element of S has the form $(x, y, -x - y)$, since $x + y + z = 0$ implies $z = -x - y$. This means that every element of S can be written uniquely as the linear combination $x(1, 0, -1) + y(0, 1, -1)$, which says the vectors $(1, 0, -1)$ and $(0, 1, -1)$ are a basis for S, and that S is two-dimensional.

Now let's look at some examples of sets with not-so-obvious geometric structure, but clear algebraic structure: solutions to homogeneous (linear) differential equations. We begin with a differential equation that every calculus student can solve: Find the general solution to the differential equation $y' - 2y = 0$. Of course the answer is $y = Ce^{2x}$ for an arbitrary constant C.

Let's say this in another way. First we observe there is a solution since it is obvious that $y = 0$ works. Now if f and g are two solutions to $y' - 2y = 0$ then so are $f \pm g$ and λf for any scalar λ. Once again, this makes the set of solutions into a vector space, this time one-dimensional, with a basis consisting of one element, $\{e^{2x}\}$.

Now let S be the set of solutions to the homogeneous differential equation $y'' + y = 0$. Depending on your calculus background, you may know the general solution, but everyone can check trivially that both $y = \cos x$ and $y = \sin x$ are solutions, and clearly $\cos x \neq \lambda \sin x$ for any choice of λ. A moment's thought shows that S is once again a vector space, so every function of the form $A \cos x + B \sin x$ is also in S. In fact every element of S has this form, so that S is a two-dimensional vector space with basis $\{\cos x, \sin x\}$.

Let's consider a final example, the set S of solutions to $y''' - 2y'' + y' - 2y = 0$. Yes, S is again a vector space, and its dimension is three. We note that $\cos x$, $\sin x$ are e^{2x} are each solutions, hence so is $Ae^{2x} + B\cos x + C\sin x$, and this is the general solution meaning that $\{e^{2x}, \cos x, \sin x\}$ is a basis for this vector space.

Consider the elliptic curve $y^2 = x^3 - 2$; more formally, put $E = \{(x, y) \mid y^2 = x^3 - 2\}$. Suppose that $P, Q \in E$, that is P and Q are points on the E. While it is easy to see that E is not a vector space (for example, λP is generally not on E, but is there some way to add P and Q to get another point on E? Indeed there is. The line through P and Q (generically) intersects the elliptic curve in a third point which we denote $P * Q$. Reflect that point across the x-axis; the resulting point is denoted $P \oplus Q$, the sum of P and Q. While it is not at all obvious, this procedure makes E into an abelian group (vector space without the scalar multiplication), and it is the structure of this group which will prove crucial for cryptography. But we have lots to do before we get there.

2. Rational Curves and Rational Points on Curves

We develop some intuition with some exercises.

Exercises.

(1) Characterize (i.e., find) all the rational points on the Fermat Curve(s) $x^n + y^n = 1$, $n > 2$.

(2) Bachet (1621) considered elliptic curves of the form $y^2 = x^3 + k$, $k \neq 0$, and proved that if (x, y) is a point on the curve, then so is $\left(\dfrac{x^4 - 8kx}{4y^2}, \dfrac{-x^6 - 20kx^3 + 8k^2}{8y^3}\right)$. While it is tedious, but straightforward to check Bachet's answer is correct, find a proof starting from our work in class. You may want to certainly want to utilize both Cartesian coordinates and some calculus, neither of which were available to Bachet!

(3) To go further we need a few definitions: We shall say that a line $ax + by + c = 0$ is called a *rational line* if $a, b, c \in \mathbb{Q}$. Similarly, we say a conic $ax^2 + bxy + cy^2 + dx + ey + f = 0$ is called a *rational conic* if $a, b, c, d, e, f \in \mathbb{Q}$.
   (a) Is every point on a rational line a rational point?
   (b) If a line passes through at least two rational points, is it a rational line? What about only one rational point?
   (c) Consider two distinct rational lines which intersect. Do they intersect in a rational point?

(4) Curves in the plane.
   (a) In how many points can two arbitrary lines (in the plane) intersect?
   (b) In how many points can a line and a conic intersect?
   (c) In how many points can two (distinct) conics intersect?
   (d) In how many points can a conic and a cubic intersect?
   (e) In how many points can two (distinct) cubics intersect?
   (f) What would be your guess for a generalization?

(5) Consider the intersection of a rational line with a rational conic.
   (a) Are the point(s) of intersection necessarily rational? (proof or counterexample).
   (b) Now let's suppose that the line intersects the conic in two points, one of which is rational. Is the second necessarily? (proof or counterexample)

2.1. A note on double roots. In the Bachet duplication formula, we made the assumption that the tangent line intersects the cubic with multiplicity two at the point of tangency. That is, if we consider the line of intersection of a tangent line at (a, b) to the elliptic curve $y^2 = f(x)$, the cubic which results will have a double root at $x = a$.

First we prove a little lemma.

LEMMA. *Let $f(x)$ be a polynomial of degree $n \geq 2$ with coefficients in a field F, and $a \in F$. Then*

(1) $f(x) = (x - a)g(x) + f(a)$.
(2) $f(a) = 0$ if and only if $f(x) = (x - a)g(x)$ for some polynomial g having coefficients in F.
(3) f has a double root at a iff $f(a) = f'(a) = 0$.

PROOF. The first item is a consequence of the division algorithm in polynomial rings, a fact you learn in 31 or 71, but probably known to you from high school. Loosely speaking when you divide one polynomial by another, you get a quotient and remainder with the remainder having degree less than the degree of the polynomial by which you divided. Thus if we divide $f(x)$ by $x - a$, we get $f(x) = (x - a)g(x) + r(x)$ for some polynomials g, r. Note that the degree of r is zero, so is a constant which we evaluate be plugging in $x = a$: $f(a) = (a - a)g(a) + r = r$.

The second statement is now immediate from the first.

So let's restrict our attention the third item. f has a double root at $x = a$ means $f(x) = (x - a)^2h(x)$ for some polynomial h. Obviously $f(a) = 0$ and the product rule for derivatives shows that $f'(a) = 0$ as well: $[f'(x) = (x - a)^2h'(x) + 2(x - a)h(x)]$.

Conversely suppose that $f(a) = f'(a) = 0$. Since $f(a) = 0$, we know that $f(x) = (x - a)g(x)$. Now $f'(x) = (x - a)g'(x) + g(x)$, so $f'(a) = 0$ means that $g(a) = 0$ which means that $g(x) = (x - a)h(x)$ for some h. Putting things together, we see that $f(x) = (x - a)g(x) = (x - a)^2h(x)$ as required. □

Now consider the curve $y^2 = f(x)$ as in the Bachet problem. Consider the tangent line to the curve at the point (a, b) $(b^2 = f(a))$. Differentiating implicitly we see that $\dfrac{dy}{dx} = \dfrac{f'(x)}{2y}$, so the slope of the tangent line at (a, b) is $m = f'(a)/2b$, and the tangent line has the form: $y - b = m(x - a)$, so substituting into $y^2 = f(x)$ we see that $f(x) = (m(x - a) + b)^2$. Let $g(x)$ be the difference: $g(x) = f(x) - (m(x - a) + b)^2$. The roots of g are the x coordinates of the points of intersection, and we claim that $g(x) = (x-a)^2(x-x_0)$, where x_0 is the x-coordinate of the point we want, but that also a is a double root.

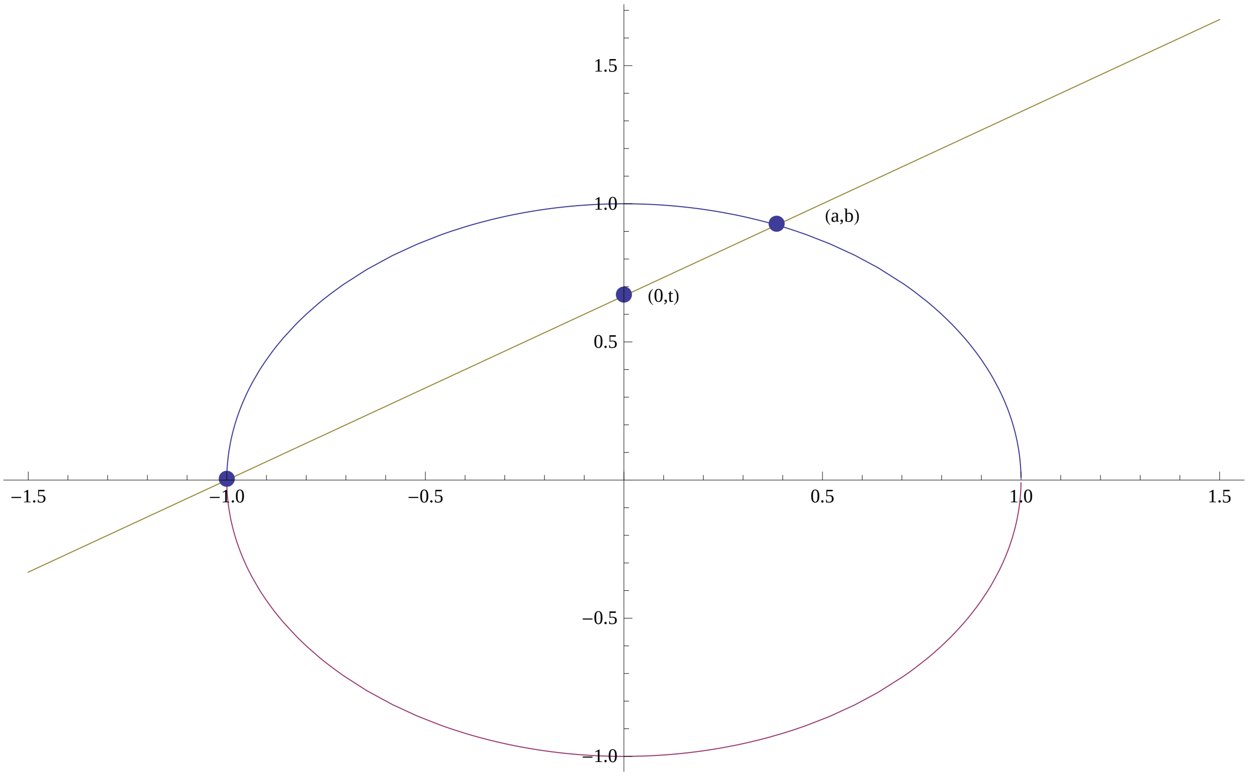
To check it is a double root, we need only check that $g(a) = g'(a) = 0$ by the lemma.

Now $g(a) = f(a) - b^2 = 0$ since (a,b) lies on the curve $y^2 = f(x)$. Next we differentiate: $g'(x) = f'(x) - 2m(m(x-a)+b)$, so that $g'(a) = f'(a) - 2mb$. But we noted that $m = f'(a)/2b$ so that $g'(a) = 0$ as well, telling us that a is a double root of g.

2.2. Parametrizing the rational points on the unit circle. With a bit of intuition gained from these exercises, we are ready to tackle a significant problem: how should we find all the rational points on a conic? Perhaps an even deeper question would be, does the set of points have any algebraic structure? Sounds like a good term project to me.

We return to our first important thread to this course: finding the rational points on $x^2+y^2=1$ is connected to finding Pythagorean triples which in turn is connected to finding congruent numbers.

We begin with a careful analysis in which we follow the arguments from [**6**]. Consider the unit circle and a point $(a,b) \neq (-1,0)$ on it as in the figure below. The line between $(-1,0)$ and (a,b) crosses the y-axis at a point $(0,t)$. Conversely, every point $(0,t)$ determines a unique point on the unit circle (except for $(-1,0)$. So there is a one-to-one correspondence between points $(0,t)$ on the y-axis and points on the unit circle (except for $(-1,0)$). Note the point $(-1,0)$ corresponds to $t = \pm\infty$.



So the line through $(-1,0)$ and $(0,t)$ (let's call it L_t) is $y = t(x+1)$. If (a,b) is the (other) point of intersection of the line and the unit circle, then $b = t(a+1)$ and $b^2 = 1 - a^2 = t^2(a+1)^2$. For a fixed value of t, $1 - a^2 = t^2(a+1)^2$ is a quadratic equation in the variable a whose roots are the x-coordinates of the points of intersection of the line with the circle. Clearly one of them is $a = -1$, so we assume $a \neq -1$, that is $1 + a \neq 0$. Now consider, $(1-a)^2 = (1-a)(1+a) = t^2(1+a)^2$ which implies (since $(1+a) \neq 0$),

$$1 - a = t^2(1+a), \text{ or}$$
$$a(t^2+1) = 1 - t^2, \text{ which yields}$$
$$a = \frac{1-t^2}{1+t^2}.$$

Substituting into the equation of the line yields $b = t(1+a) = \dfrac{2t}{1+t^2}$

So every point on the unit circle (except $(-1,0)$) has the form $(a,b) = \left(\dfrac{1-t^2}{1+t^2}, \dfrac{2t}{1+t^2}\right)$, for some value of $t \in \mathbb{R}$. Notice that

$$\lim_{t\to\infty}\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right) = \lim_{t\to\infty}\left(\frac{1/t^2-1}{1/t^2+1}, \frac{2}{1/t+t}\right) = (-1,0).$$

From our earlier observations, if (a, b) is a rational point, then the line L_t is a rational line. Since the y-axis ($x = 0$) is also a rational line, their point of intersection, $(0, t)$, is a rational point, hence t is rational. Conversely, if t is rational, it is obvious from the formulas above that (a, b) is a rational point on the circle. So there is a one-to-one correspondence between the rational points on the circle (except for $(-1, 0)$) and rational values of t.

2.3. Finding all Pythagorean triples. We now use the above formulas to find all Pythagorean triples: positive integers A, B, C with $A^2 + B^2 = C^2$. First we simplify the problem a bit. In enumerating the triples, there is no reason to consider triples with a common divisor. For example if A, B, C have a greatest common divisor t, then $A = A_0 t$, $B = B_0 t$, and $C = C_0 t$ with A_0, B_0, C_0 being relatively prime (no common divisor other than one), and A_0, B_0, C_0 is a Pythagorean triple from which we could recover the original by multiplying the side lengths all by t. We call a Pythagorean triple *primitive* if the three sides of the right triangle have relatively prime lengths.

Actually, we note that if $A^2 + B^2 = C^2$ and A, B, C have no common divisor other than one, then the pairs $\{A, B\}$, $\{A, C\}$, and $\{B, C\}$ are also relatively prime. For example, if $t \mid B$ and $t \mid C$, then $t^2 \mid (C^2 - B^2) = A^2$ which means $t \mid A$.

Thus starting with a primitive Pythagorean triple satisfying $A^2 + B^2 = C^2$, we produce a rational point $(A/C, B/C)$ on the unit circle in which the rational numbers A/C, B/C are already in lowest terms. But we have formulas for the rational points on the unit circle in terms of the rational number t which we write as $t = m/n$ with m, n relatively prime positive integers, Substituting into our previous formula says that

$$\frac{A}{C} = \frac{n^2 - m^2}{n^2 + m^2} \qquad \text{and} \qquad \frac{B}{C} = \frac{2mn}{n^2 + m^2}.$$

It is easy to see that A, B must have different parity (i.e., one odd, one even): Clearly they both can't be even since their gcd is one. They also can't be odd. In terms of congruences, if A and B were both odd, then $A^2, B^2 \equiv 1 \pmod 4$, so $A^2 + B^2 \equiv 2 \pmod 4$. This means C must be even so that $C^2 \equiv 0 \pmod 4$. This provides a contradiction since if $A^2 + B^2 = C^2$ then surely this must hold as a congruence modulo 4, but $2 \not\equiv 0 \pmod 4$. So without loss of generality (since we can't tell the two legs of the right triangle apart!), we shall assume that A is odd and B is even.

Before continuing we make a few technical observations. Since $A, B, C > 0$, the point $(A/C, B/C)$ is in the first quadrant and not equal to $(0, 1)$ or $(1, 0)$ which means the corresponding parameter t satisfies $0 < t < 1$. Since $t = m/n$ we may assume that $n > m > 0$.

Now since A/C and B/C are already in lowest terms, we must have positive integers (here we use that $n > m$) λ, μ so that

$$\begin{aligned} n^2 - m^2 &= \lambda A \qquad & 2mn \qquad &= \mu B \\ n^2 + m^2 &= \lambda C \qquad & n^2 + m^2 &= \mu C. \end{aligned}$$

Since $C \neq 0$, $\mu C = \lambda C$ implies $\mu = \lambda$, so there is a positive integer λ with

$$\lambda A = n^2 - m^2, \qquad \lambda B = 2mn, \qquad \lambda C = n^2 + m^2.$$

The claim is that $\lambda = 1$ so that our Pythagorean triples are already determined.

We note that $\lambda(A + C) = 2n^2$ while $\lambda(C - A) = 2m^2$. This means that λ divides both $2m^2$ and $2n^2$, but m and n are relatively prime which means λ divides 2, so either $\lambda = 1$, or 2. We need only exclude $\lambda = 2$ as possible.

Recall that we are assuming that A is odd and B is even, so if $\lambda = 2$, $\lambda A \equiv 2 \pmod 4$ while $n^2 - m^2 \equiv 0, 1, 3 \pmod 4$, a contradiction. So we must have $\lambda = 1$.

Theorem 2.1. *Every primitive Pythagorean triple A, B, C with B even has the form*

$$A = n^2 - m^2, \qquad B = 2mn, \qquad C = n^2 + m^2$$

for relatively prime positive integers m, n.

Note that we necessarily have $n > m$. For triples with A even, we could interchange formulas.

2.4. Implementing the algorithm. We end with some pseudocode (actually valid Mathematica code) to list congruent numbers associated to Pythagorean triples for $1 \geq m < n < 6$. The first line defines a function which extracts the square-fre part of an integer.

```
SquarefreePart[n_Integer?Positive] :=
 Times @@ Power @@@ ({#[[1]], Mod[#[[2]], 2]} & /@ FactorInteger[n])

For [n = 2, n < 6, n++,
 For[m = 1, m < n, m++,
  If[GCD[m, n] == 1,
   Print["(m,n) = (", m, ",", n, "), Pythagorean Triple = (",
    n^2 - m^2, ",", 2 m*n, ",", n^2 + m^2, "), Congruent number = ",
    SquarefreePart[m*n*(n^2 - m^2)]]]]]
```

Output: Note that the last line of output recovers that 5 is a congruent number.

```
(m,n) = (1,2), Pythagorean Triple = (3,4,5), Congruent number = 6

(m,n) = (1,3), Pythagorean Triple = (8,6,10), Congruent number = 6

(m,n) = (2,3), Pythagorean Triple = (5,12,13), Congruent number = 30

(m,n) = (1,4), Pythagorean Triple = (15,8,17), Congruent number = 15

(m,n) = (3,4), Pythagorean Triple = (7,24,25), Congruent number = 21

(m,n) = (1,5), Pythagorean Triple = (24,10,26), Congruent number = 30

(m,n) = (2,5), Pythagorean Triple = (21,20,29), Congruent number = 210

(m,n) = (3,5), Pythagorean Triple = (16,30,34), Congruent number = 15

(m,n) = (4,5), Pythagorean Triple = (9,40,41), Congruent number = 5
```

CHAPTER 3

Some Elementary Number Theory

Most of this section is very standard, see e.g., [**2**] or [**5**].

1. Basic Properties of the Integers

Let $\mathbb{N} = \{0, 1, 2, \dots\}$ denote the natural numbers. By construction, the natural numbers are *well-ordered*, that is every nonempty subset of $\mathbb{N}$ contains a least element. This is a fundamental fact in number theory and is equivalent to both notions of mathematical induction.

In elementary school, you learn to do long division of integers, and can show for example that 257 divided by 12 has quotient 21 and remainder 5. Put another way, $257 = 12(21) + 5$. There are many things we assumed about this process. One was that the remainder was nonnegative and smaller than the number by which you were dividing, and the second is that these numbers are unique. This is summarized by the following theorem:

Theorem. *Let $a, b \in \mathbb{Z}$, $b > 0$. Then there exist unique integers q and r with $a = bq + r$ and $0 \le r < b$.*

Remark. Intuitively, the existence part is easy to see: Assume for convenience that $a > 0$. Then a lies in some interval $[qb, (q+1)b)$, so $a - qb = r$ lies in $[0, b)$. Formally, we have the proof:

Proof. First we show existence. Let $S\{a + nb \mid n \in \mathbb{Z}\} = \{a, a \pm b, a \pm 2b, \dots\}$, and consider $S \cap \mathbb{N}$. We see that $a - nb \ge 0$ if and only if $n \le -a/b$, so certainly $S \cap \mathbb{N}$ is nonempty, and so by well-ordering, has a least element $r = a - qb$. All we need to verify is that $r < b$. If not, $r \ge b$, so $r - b \ge 0$ and $r - b = a - b(q+1) \in S \cap \mathbb{N}$ and is strictly smaller than r, contradicting that r is the smallest element of $S \cap \mathbb{N}$. Thus $a = bq + r$ with $0 \le r < b$.

As to uniqueness, if $a = bq + r = bq' + r'$ with $0 \le r' < b$, then we derive that $b(q - q') = r' - r$. Now $0 \le r, r' < b$ so their difference must be less than b, but is a multiple of b, hence zero. Thus $r = r'$ and since then $b(q - q') = 0$ we have $q = q'$ as well. □

When $a = bq + r$ and $r = 0$ we say that b divides a, written $b \mid a$, so for example $2 \mid 26$ while $3 \nmid 35$.

Example:

- Show that $3 \mid 0$, but $0 \nmid 3$.
- Show that if $a \mid b$ and $b \mid c$, then $a \mid c$.
- Show that if $a \mid b$ and $c \mid d$, then $ac \mid bd$.
- Show that if $m \neq 0$, then $a \mid b$ if and only if $am \mid bm$.

Proposition. *(1) If $a \mid b_1, \ldots, b_r$, then $a \mid m_1 b_1 + \ldots m_r b_r$ for any integers m_i.*
(2) If $a \mid b$ and $b \mid a$, then $a = \pm b$.

Definition. If $d \mid a$ and $d \mid b$, we say d is a *common divisor* of a and b. If $d = \pm 1$ are the only common divisors of a and b, we say that a, b are *relatively prime* or *coprime*.

Remark. Often in elementary number theory, we focus more on the positive integers, so would say $1, 2, 3, 6.$ 12 are the common divisors of $24, 36$, when indeed the more correct statement would be that $\pm 1, \pm 2, \pm 3, \pm 4, \pm 6$, and ± 12 are the common divisors of 24 and 36, while $24, 25$ are relatively prime.

Definition. Let $a, b \in \mathbb{Z}$, not both zero. The *greatest common divisor* of a, b is the unique positive integer d so that

(1) $d \mid a$ and $d \mid b$ (i.e., d is a common divisor), and
(2) If $c \mid a$ and $c \mid b$, then $c \mid d$ (so d is greatest among positive divisors).

We denote this as $d = \gcd(a, b)$ or simply $d = (a, b)$ if the context is clear.

Example: $\gcd(12, 15) = 3$, $\gcd(24, 36) = 12$, $\gcd(24, 25) = 1$.

For small integers, we tend to rely on our ability to factor integers; for large numbers factoring is impractical (indeed the security of RSA encryption depends upon that fact), and we need to rely on a more computationally feasible means of extracting the gcd, Euclid's algorithm.

Lemma. *If a, b are integers, not both zero and with $a = bq + r$, then $\gcd(a, b) = \gcd(b, r)$.*

Proof. We claim that the set of common divisors of a and b are the same as the set of common divisors of b and r. Given this claim, the greatest among these would necessarily be the same. To establish the claim we need only show that every common divisor of a, b is a common divisor of b, r, and conversely.

If $d \mid a, b$, then by a recent proposition d divides any linear combination of them, namely $d \mid r = a - bq$. Conversely, if $d \mid b, r$ then by the same reasoning $d \mid a = bq + r$. □

A few easy observations about gcds.

Proposition. *Suppose that $a, b \in \mathbb{Z}$, not both zero. Then*

- $\gcd(\pm a, \pm b) = \gcd(|a|, |b|)$.
- *If $a \neq 0$, the* $\gcd(a, 0) = |a|$.
- *If $a \neq 0$, then* $\gcd(a, a) = |a|$.
- *If $b \mid a$, then* $\gcd(a, b) = |b|$.

2. Euclid's Algorithm

We give a slightly restricted version of Euclid's algorithm. Assume $a \geq b > 0$ are integers and we wish to find their gcd. We iterate the division algorithm as follows:

$$\begin{aligned}
a &= bq_1 + r_1, & 0 &\leq r_1 < b \\
b &= r_1 q_2 + r_2, & 0 &\leq r_2 < r_1 \\
r_1 &= r_2 q_3 + r_3, & 0 &\leq r_3 < r_2 \\
&\vdots \\
r_{n-3} &= r_{n-2} q_{n-1} + r_{n-1}, & 0 &\leq r_{n-1} < r_{n-2} \\
r_{n-2} &= r_{n-1} q_n + r_n, & r_n &= 0
\end{aligned}$$

Note that $0 \leq r_n < r_{n-1} < \cdots < r_1 < b$ is a strictly decreasing sequence of non-negative numbers, so the algorithm must terminate in fewer than $|b|$ steps, generally much faster.

The point of Euclid's Algorithm is that

Theorem. *With the notation as above,* $\gcd(a, b) = r_{n-1}$*, that is the last nonzero remainder in Euclid's algorithm.*

Proof. By the lemma, $\gcd(a, b) = \gcd(b, r_1) = \gcd(r_1, r_2) = \cdots = \gcd(r_{n-2}, r_{n-1}) = \gcd(r_{n-1}, 0) = r_{n-1}$. □

Example 2.1. Let's compute the gcd of 252 and 198.

$$\begin{aligned}
252 &= 198(1) + 54 \\
198 &= 54(3) + 36 \\
54 &= 36(1) + 18 \\
36 &= 18(2) + 0
\end{aligned}$$

So the $\gcd(252, 198) = 18$, the last nonzero remainder.

Now we come to a major application of Euclid's algorithm.

Theorem. *(Bezout) Let a, b be integers, not both zero. Then there exists $u, v \in \mathbb{Z}$ so that $gcd(a, b) = au + bv$.*

Proof. Note that there is no loss of generality assuming both a and b are non-negative, for say $a < 0$ and $b \geq 0$. We have $\gcd(a,b) = \gcd(|a|,b) = \gcd(-a,b) = -a(u) + bv = a(-u) + bv$ as desired.

The proof is simply to realize the we can run Euclid's algorithm backwards starting with the gcd which equals r_{n-1} and back substituting until we have a combination of a and b. We illustrate this with the example we computed above: $\gcd(252,198) = 18$.

$$\begin{aligned} 18 &= 54 - 36(1) \\ &= 54 - (1)(198 - 54(3)) = 198(-1) + 54(4) \\ &= 198(-1) + 4(252 - 198(1)) = 198(-5) + 252(4). \end{aligned}$$

□

From Bezout's theorem, we obtain two hugely useful corollaries.

Corollary. *Let $a, b \in \mathbb{Z}$ not both zero. Then*

(1) If $a \mid c$, $b \mid c$, and $\gcd(a,b) = 1$, then $ab \mid c$.
(2) If $a \mid bc$ and $\gcd(a,b) = 1$, then $a \mid c$.

Proof. Since the $\gcd(a,b) = 1$, Bezout's theorem says there are integers u, v so that $au + bv = 1$. For the first assertion, write $c = am = bn$. Then

$$c = c \cdot 1 = c(au + bv) = cau + cbv = (bn)au + (am)bv = ab(nu + mv),$$

showing that $ab \mid c$.

Similarly, for the second, we write

$$c = c \cdot 1 = c(au + bv) = cau + cbv.$$

We note that $a \mid a$ and $a \mid bc$ by hypothesis, so a divides the linear combination $cau + cbv = c$. □

3. Modular Arithmetic, Equivalence Relations

Let n be a positive integer, and a, b arbitrary integers. Write $a = nq + r$ and $b = nq' + r'$ with $0 \leq r, r' < n$ as in the division algorithm.

Definition. We say that a is *congruent to b modulo n*, written $a \equiv b \pmod{n}$ or $a \equiv b(n)$ if and only if $r = r'$. Equivalently, $a \equiv b \pmod{n}$ if and only if $n \mid (b - a)$, which is also equivalent to saying that $a = b + kn$ for some integer k.

For example $a \equiv b \pmod{2}$ means $2 \mid (b-a)$, or that a and b have the same parity (are both odd or both even).

It is easy to check that congruence modulo n is an equivalence relation (defined below).

As interim notation, to distinguish clearly between equivalence classes of integers from the integers which comprise the class, we shall write

$$[a]_n = \{k \in \mathbb{Z} \mid k \equiv a \pmod{n}\} = \{a + \ell n \mid \ell \in \mathbb{Z}\}.$$

So for example,

$$\begin{aligned}[0]_2 &= \{0, \pm 2, \pm 4, \pm 6, \dots\} = [2]_2 = [-2048]_2 \text{ (the even integers)}\\ [1]_2 &= \{\pm 1, \pm 3, \pm 5, \dots\} = [3]_2 = [-12345]_2 \text{ (the odd integers),}\end{aligned}$$

and the congruence classes form a *partition* of $\mathbb{Z}$: $\mathbb{Z} = [0]_2 \cup [1]_2$ and $[0]_2 \cap [1]_2 = \emptyset$.

If we look at congruence modulo 5, we see that there are five congruence classes $[0]_5, [1]_5, [2]_5, [3]_5, [4]_5$. Every integer lies in one of these classes since every integer $a = 5q + r$ for a uniquely determined r, $0 \le r \le 4$, and by the uniqueness, it can only lie in one class. If we let $\mathbb{Z}_n = \{[a]_n \mid a \in \mathbb{Z}\}$, then $\mathbb{Z}_2 = \{[0]_2, [1]_2\}$, $\mathbb{Z}_5 = \{[0]_5, [1]_5, [2]_5, [3]_5, [4]_5\}$, and in general $\mathbb{Z}_n = \{[0]_n, [1]_n, \dots, [n-1]_n\}$. We have that $\mathbb{Z} = [0]_n \cup [1]_n \cup \dots \cup [n-1]_n$ (which says that every integer lies in a congruence class, and because the division algorithm gives unique remainders, an integer can lie in only one congruence class. It follows that these classes are disjoint and so once again form a partition of $\mathbb{Z}$.

It is actually not difficult to show that there is a one-to-one correspondence between partitions of a set and equivalence relations on the set, so it seems we should at least define an equivalence relation.

Let S be a set. Formally (so this is going to sound really technical for a few seconds), a relation on a set S is a subset $R \subseteq S \times S$, that is to say some collection of ordered pairs (a, b), where $(a, b) \in R$ means a is related to b. For example, suppose that $S = \mathbb{Z}$ are the relation we want to talk about is the notion of "less than". So $2 < 3$, but $3 \not< 2$. So our relation $R \subset \mathbb{Z} \times \mathbb{Z}$ contains the ordered pair $(2, 3)$ (since $2 < 3$), but does not contain the ordered pair $(3, 2)$.

It is clear even from this simple example that a relation need not be symmetric (that is a related to b does not imply that b is related to a). It is also clear, that for a general relation an element a need not be related to itself since $a \not< a$. The relations we wish to consider are nicer.

We will write as short hand $a \sim b$ if a is related to b (or more formally that $(a, b) \in R$). The relation we want to consider is that of congruence modulo a positive integer n.

So we shall write $a \sim b$ if and only if $a \equiv b \pmod{n}$ (which is to say $n \mid (a-b)$). We see that congruence is an equivalence relation since it satisfies the following three properties:

(1) $a \equiv a \pmod{n}$, for all $a \in \mathbb{Z}$. (reflexive)
(2) $a \equiv b \pmod{n}$ implies $b \equiv a \pmod{n}$, for all $a, b \in \mathbb{Z}$ (symmetric), and
(3) $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ implies $a \equiv c \pmod{n}$, for all $a, b, c \in \mathbb{Z}$ (transitive).

These are easily verified: the first since $n \mid (a-a) = 0$; the second since $n \mid (a-b)$ if and only if $n \mid (b-a)$; and the third since if $n \mid (a-b)$ and $n \mid (b-c))$, then $n \mid (a-b)+(b-c) = (a-c)$.

Now we need to be able to do some arithmetic with congruences which will eventually translate to arithmetic on the congruence classes in $\mathbb{Z}_n$.

Proposition. *Let $a, b, a', b' \in \mathbb{Z}$, n a positive integer with $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$. Then*

(1) $a \pm b \equiv a' \pm b' \pmod{n}$.
(2) $ab \equiv a'b' \pmod{n}$.

Proof. $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$ means that $a = a' + kn$ and $b = b' + \ell n$ for some integers k and ℓ. Thus,

$$\begin{aligned} a \pm b &= a' \pm b' + n(k \pm \ell), \\ ab &= a'b' + n(kb' + \ell a' + k\ell n). \end{aligned}$$

Rewriting these equalities as congruences yields the result. □

Example 3.1.

$$10045 \cdot 19 \equiv 5 \cdot 19 \equiv 5 \cdot (-1) \equiv -5 \equiv 15 \pmod{20}.$$

Example 3.2. Is 12345678 a square in $\mathbb{Z}$?

Well, integers are either odd or even, that is of the form $m = 2n$ or $m = 2n+1$, so their squares $m^2 = 4n^2 or 4n^2 + 4n + 1$ satisfy $m^2 \equiv 0, 1 \pmod{4}$. That is any integer which is congruent to 2 or $w \pmod{4}$ cannot be a square. And indeed we see (using the base 10 expansion of our number and the fact that $100 \equiv 0 \pmod{4}$) that $12345678 = 123456(100) + 78 \equiv 78 \pmod{4} \equiv 2 \pmod{4}$, so indeed 12345678 is not a square.

Example 3.3. Find the last decimal digit of $1! + 2! + \cdots + 237!$. A moment's thought tells us that given the base 10 expansion of an integer, the last digit is the least non-negative residue modulo 10. Moreover, we also note that for $n \geq 5$, $n! \equiv 0 \pmod{10}$, so $1! + 2! + \cdots + 237! \equiv 1! + 2! + 3! + 4! \equiv 1 + 2 + 6 + 24 \equiv 33 \equiv 3 \pmod{10}$.

Example 3.4. Finally, when we talked about Pythagorean triples: integers A, B, C with $A^2 + B^2 = C^2$, we asserted and awkwardly justified that A and B could not both be odd. Now we easily see why. A square integer is congruent to 0 or 1 modulo 4, so $C^2 \equiv 0, 1 \pmod{4}$. If A and B are both odd, that $A^2 + B^2 \equiv 2 \pmod{4}$, so there can be no equality $A^2 + B^2 = C^2$.

4. Elementary Cryptography: Caesar Cipher

We'll leave historical background about cryptography as an exercise in prelude to writing term papers, but the Caesar cipher was a very early example of a cryptographic system with mathematical underpinnings.

To send private messages (say) to his commanders on the battlefield, Caesar would encrypt them by writing out the message, and then shifting each letter in the message forward by three: That is:

Plaintext:	AB	CD	EF	GH	IJ	KL	MN	OP	QR	ST	UV	WX	YZ
	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$	$\updownarrow\updownarrow$
Ciphertext:	DE	FG	HI	JK	LM	NO	PQ	RS	TU	VW	XY	ZA	BC

Thus the word CAT in plaintext would translate to FDW in ciphertext.

Mathematically, we achieve this as follows:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

So C-A-T would be encoded 2-0-19; the ciphertext F-D-W would be 5-3-22. So that if P is the numeric equivalent of a plaintext letter, the corresponding ciphertext letter would be determined by the congruence

$$C \equiv P + 3 \pmod{26}.$$

The plaintext would be recovered from the ciphertext by

$$P \equiv C - 3 \equiv C + 23 \pmod{26}.$$

The Caesar cipher is an example of a *shift* cipher, i.e. of the form $C \equiv P + k \pmod{n}$. The values of k which produce distinct (nontrivial) ciphers is called the *keyspace* associated to the encryption scheme; clearly, it has size $n - 1$ ($k = 0$ is trivial encryption).

Any person trying to break this cryptosystem would have little challenge, especially if they suspected the nature of the encryption scheme.

A basic cryptographic set up has two functions E and D, representing an encryption and decryption scheme. If P denotes a plaintext message, and C the corresponding ciphertext (encrypted message), then the requirements for a cryptosystem are pretty basic.

We take plaintext P and use E to encrypt the message, producing ciphertext $C = E(P)$. Ideally, it is very difficult to discover P from C. The other essential feature is that the decryption scheme must work, that is $P = D(C) = D(E(P))$, D is a left-hand inverse to E.

For the Caesar cipher we had E and D defined as:

$$\begin{aligned} C = E(P) &\equiv P + 3 \pmod{26} \\ P = D(C) &\equiv C - 3 \pmod{26} \equiv C + 23 \pmod{26} \end{aligned}$$

Let's consider a slightly more complicated scheme: affine ciphers. Here we take $C \equiv aP + b \pmod{26}$ for integers a and b. Clearly $a = 1$ recovers our shift cipher, but when $a \neq 1$, we must ask when is the congruence $C \equiv aP + b \pmod{26}$ uniquely solvable for $P \pmod{26}$.

In class exercises

Let's summarized what we have learned from the exercises. We begin with a dimple case:

Proposition. *The congruence $ax \equiv 1 \pmod{n}$ is solvable if and only if $\gcd(a, n) = 1$, and when solvable, there is a unique solution modulo n.*

Proof. If $ax \equiv 1 \pmod{n}$ is solvable, then there exists a $y \in \mathbb{Z}$ with $ax + ny = 1$. Let d be the gcd of a and n. Then, by definition of a gcd, $d \mid a$ and $d \mid n$, so we know that d divides any combination of a and n; in particular $d \mid (ax + ny) = 1$, so d is clearly one.

Conversely, if $d = \gcd(a, n) = 1$, then Bezout's theorem tells us that there exists $u, v \in \mathbb{Z}$ with $au + nv = d = 1$, but this means $au \equiv 1 \pmod{n}$, so the congruence is solvable.

To see uniqueness, if $ax \equiv ay \equiv 1 \pmod{n}$, then $uax \equiv uay \pmod{n}$. But as $au \equiv 1 \pmod{n}$, we deduce $x \equiv y \pmod{n}$ as required. □

Implicit in the proof above is the following result allowing us to recognize when two integers are relatively prime.

Corollary. *Let $a, b \in \mathbb{Z}$. Then $\gcd(a, b) = 1$ if and only if there exist $u, v \in \mathbb{Z}$ with $au + bv = 1$.*

Now we proceed to handle general linear congruences.

Proposition. *If the congruence $ax \equiv b \pmod{n}$ is solvable, then $d = \gcd(a, n) \mid b$.*

Note that the contrapositive is more instructive: If $d \nmid b$, then the congruence is not solvable.

Proof. The proof is similar to the one above. If $ax \equiv b \pmod{n}$ is solvable, then there exist integers x, y so that $ax + ny = b$. Since $d = \gcd(a, n)$, we know $d \mid a$ and $d \mid n$, so d divides any combination of a and n, in particular, $d \mid b$. □

The converse is where the substance lies.

THEOREM. *Let $d \mid \gcd(a,n)$. If $d \mid b$, then the congruence $ax \equiv b \pmod{n}$ is solvable, and there are precisely d incongruent solutions modulo n. Indeed that are all of the form $x_0 + \frac{n}{d}t$ for $t = 0, 1, \ldots, d-1$ for any particular solution x_0.*

Before proving the theorem, we sum up the results so far as

COROLLARY. *The congruence $ax \equiv b \pmod{n}$ is solvable if and only if $d = \gcd(a,n) \mid b$. When solvable, there are precisely d incongruent solutions modulo n.*

OF THEOREM. Let $d = \gcd(a,n)$ and assume that $d \mid b$. By Bezout, we know there exists integers u, v so that

$$\begin{aligned} au + nv &= d \\ \frac{a}{d}u + \frac{n}{d}v &= 1 \qquad \text{(divide by } d\text{)} \\ \frac{a}{d}ub + \frac{n}{d}vb &= b \qquad \text{(multiply by } b\text{)} \\ a(u\frac{b}{d}) + n(v\frac{b}{d}) &= b \qquad \text{(redistribute since } d \mid b\text{)} \end{aligned}$$
$$.$$

This says that $x_0 = u\frac{b}{d}$ is a solution to $ax \equiv b \pmod{d}$, so solutions exist. Now, how many are there. Suppose that y_0 is another solution to the congruence. Then

$$\begin{aligned} ay_0 \equiv ax_0 \equiv b \pmod{n} &\iff a(y_0 - x_0) \equiv 0 \pmod{n} \\ &\iff a(y_0 - x_0) = nk \quad \text{for some integer } k \\ &\iff \frac{a}{d}(y_0 - x_0) = \frac{n}{d}k \\ &\iff \frac{a}{d}(y_0 - x_0) \equiv 0 \pmod{\frac{n}{d}} \\ &\Rightarrow u\frac{a}{d}(y_0 - x_0) \equiv 0 \pmod{\frac{n}{d}} \\ &\Rightarrow (y_0 - x_0) \equiv 0 \pmod{\frac{n}{d}} \\ &\Rightarrow y_0 = x_0 + \frac{n}{d}t \quad \text{for } t = 0, 1, \ldots, d-1, \end{aligned}$$

all of which are distinct modulo n. □

Note that $ax \equiv b \pmod{n}$ is solvable if and only if $\frac{a}{d}x \equiv \frac{b}{d} \pmod{\frac{n}{d}}$, and modulo $\frac{n}{d}$ there is a unique solution which propagates back to d solutions modulo n. We consider some examples.

EXAMPLE 4.1. We consider $6x \equiv 9 \pmod{15}$. We see that $d = \gcd(6, 15) = 3$ divides 9, so there will be 3 incongruent solutions modulo 15. As above we see that $6x \equiv 9$

(mod 15) $\iff$ $2x \equiv 3 \pmod 5$, to which there is a unique solution $x \equiv 4 \pmod 5$. Thus modulo 15, the solutions are $x \equiv 4 + 5t \pmod{15}$, $t = 0, 1, 2$, that is $x \equiv 4, 9, 14 \pmod{15}$.

Next we do a more complicated example which mimics the proof and uses Bezout's theorem.

EXAMPLE 4.2. We consider $198x \equiv 90 \pmod{252}$. We begin by computing the gcd of 252 and 198 (which we have done in an earlier class).

$$\begin{aligned} 252 &= 198(1) + 54 \\ 198 &= 54(3) + 36 \\ 54 &= 36(1) + 18 \\ 36 &= 18(2) + 0 \end{aligned}$$

So $d = \gcd(252, 198) = 18$, and we see that $18 \mid 90$ so the congruence is solvable and has 18 incongruent solutions modulo 252.

To gain a particular solution, we work Euclid's algorithm backwards, solving for 18 as follows:

$$\begin{aligned} 18 &= 54 - 36(1) \\ &= 54 - (1)(198 - 54(3)) = 198(-1) + 54(4) \\ &= 198(-1) + 4(252 - 198(1)) = 198(-5) + 252(4). \end{aligned}$$

Now

$$\begin{aligned} 198(-5) + 252(4) &= 18 \quad \text{means} \\ 198(-25) + 252(20) &= 90 \quad \text{(multiply by 5)} \end{aligned}$$

so $x_0 = -25$ is one solution to the original congruence. All (modulo 252) are of the form $x_0 + \frac{n}{d}t = -25 + 14t$, $t = 0, 1, \ldots, 17$.

Note that alternatively, we might have simplified the congruence $198x \equiv 90 \pmod{252}$ to the equivalent $11x \equiv 5 \pmod{14}$, a quick inspection of which suggests $x \equiv 3 \pmod{14}$ as a solution. Thus the solutions to the original congruence are of the form $3+14t$, $t = 0, 1, \ldots, 17$.

While these may look rather different, they are not. The first set of solutions produces the congruence classes:

$$[-25]_{252}, [-11]_{252}, [3]_{252}, [17]_{252}, \ldots, [213]_{252},$$

while the second set produces

$$[3]_{252}, [17]_{252}, \ldots, [213]_{252}, [227]_{252}, [241]_{252}.$$

The classes $[3]_{252} - [213]_{252}$ are obviously the same, and we see $[-25]_{252} = [227]_{252}$ and $[-11]_{252} = [241]_{252}$.

As a final example, we consider a simple congruence with solutions obtained in three different ways.

Example 4.3. The congruence $10x \equiv 15 \pmod{35}$ is solvable since $d = \gcd(10, 35) = 5 \mid 15$, and there will be 5 solutions modulo 35.

We can always reduce to the congruence $2x \equiv 3 \pmod 7$.

Solution 1 is by inspection: $x \equiv 4 \pmod 7$ works and is the unique solution mod 7, so we get $x \equiv 4 + 7t \pmod{35}$, $t = 0, 1, 2, 3, 4$ for the 5 solutions modulo 35.

Solution 2 says $2x \equiv 3 \pmod 7$ is solvable since $gcd(2, 7) = 1$. Use Bezout to write $1 = 2(-3) + 7(1)$. Multiplying by 3 gives $2(-9) + 7(3) = 3$, that is $x \equiv -9 \pmod{35}$ is a solution to the original congruence, so solutions are $-9 + 7t \pmod{35}$, $t = 0, 1, 2, 3, 4$.

Solution 3 says maybe you can guess a solution to $2x + 7y = 3$, say $2(-2) + 7(1) = 3$. This would give $x \equiv -2 + 7t \pmod{25}$, $t = 0, 1, 2, 3, 4$ for a complete set of solutions.

CHAPTER 4

A bit of group theory: $\mathbb{Z}_n$ and U_n as groups

When n is understood, we shall write $\mathbb{Z}_n = \{[0]_m, \dots, [n-1]_n\}$ as $\mathbb{Z}_n = \{\overline{0}, \overline{1}, \dots, \overline{n-1}\}$. And we shall typically use one of two sets of representatives of the congruence classes, least non-negative residues, or least absolute residues (least in absolute value). For example,

$$\mathbb{Z}_6 = \{\overline{0}, \overline{1}, \dots, \overline{5}\} = \{-\overline{2}, -\overline{1}, \overline{0}, \overline{1}, \overline{2}, \overline{3}\}$$
$$\mathbb{Z}_5 = \{\overline{0}, \overline{1}, \dots, \overline{4}\} = \{-\overline{2}, -\overline{1}, \overline{0}, \overline{1}, \overline{2}\}$$

Consider $\mathbb{Z}_n$, the set of n congruence classes modulo n. We wish to define operations on this set which will make it into what is known as a ring.

We define:

$$[a]_n + [b]_n = [a+b]_n \quad (\overline{a} + \overline{b} = \overline{a+b})$$
$$[a]_n - [b]_n = [a-b]_n \quad (\overline{a} - \overline{b} = \overline{a-b})$$
$$[a]_n * [b]_n = [a*b]_n \quad (\overline{a} * \overline{b} = \overline{a*b})$$

When one does something like this where the elements themselves depend on how they are named, we must check that things are *well-defined.*

For example, if $[a]_n = [a']_n$ and $[b]_n = [b']_n$, there could be an ambiguity of answer since our rule says $[a]_n + [b]_n = [a+b]_n$, but it must be the same as $[a']_n + [b']_n = [a'+b']_n$. But we have already checked this via congruences. In that context it would say that if $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$, then $a + b \equiv a' + b' \pmod{n}$, the same being true if we replace $+$ with $-$ or $*$.

So we now have a set with well-defined operations on it. We now investigate some of those properties, but first we begin with a more general context.

Definition. A nonempty set G is called a group if it has a binary operation * (that is, a map $G \times G \to G$ written $(g,h) \mapsto g * h$) satisfying

- (identity) There is an element $e \in G$ so that $g * e = e * g = g$ for all elements $g \in G$.
- (inverses) For every $g \in G$ there is an $h \in G$ so that $g * h = h * g = e$.
- (associative) For all $g, h, k \in G$, $g * (h * k) = (g * h) * k$.

The group is called abelian (commutative) if the operation is commutative, that is $g*h = h*g$ for all $g, h \in G$.

Example 0.4.

- $(\mathbb{Z}, +)$, the integers under addition is an abelian group. The identity is 0, and the inverse of m is $-m$.
- $(\mathbb{Z}_n, +)$, the integers modulo n is an abelian group under addition. The identity is $[0]_n$ and the inverse of $[a]_n$ is $[-a]_n$
- $(\mathbb{Z}_n, \cdot)$ is not a group under multiplication. The element $[0]_n$ has no multiplicative inverse, though $[1]_n$ acts as an identity.
- $U_n = \{[a]_n \mid \gcd(a, n) = 1\}$ is an abelian group under multiplication. First we should prove it is closed under multiplication, that is if $[a]_n, [b]_m \in U_n$, then so is $[a]_n[b]_n = [ab]_n$. This boils down to saying that if $\gcd(a, n) = 1 = \gcd(b, n)$, then $\gcd(ab, n) = 1$. If $\gcd(a, n) = 1 = \gcd(b, n)$, then by Bezout, there exist integers u, v, u', v' so that $au + nv = 1 = bu' + nv'$. Multiplying them together shows that $ab(uu') + n(bu'v + auv' + nvv') = 1$ which implies ab and n are coprime.

  Given closure, the identity is $[1]_n$ and an element $[a]_n$ has a inverse since $[a]_n[b]_n = [1]_n = [b]_n[a]_n$ is solvable for b iff the congruence $ax \equiv 1 \pmod{n}$ is solvable. We know it has a unique solution modulo n since $\gcd(a, n) \mid 1$.

We add a final example noting that $(\mathbb{Z}_n, +, \cdot)$ is a set with two well-defined operations. It is an abelian group under $+$, has an identity under $\cdot$ which is associative, and satisfies two distributive laws: $[a]([b] + [c]) = [a][b] + [a][c]$ and $([a] + [b])[c] = [a][c] + [b][c]$. This makes $(\mathbb{Z}_n, +, \cdot)$ into a commutative ring. The sets $(\mathbb{Q}, +, \cdot)$ and $(\mathbb{Z}, +, \cdot)$ are also rings.

Definition. We define the Euler phi function by

$$\phi(n) = \#U_n = \#\{k \mid 1 \leq k \leq n, \quad \gcd(k, n) = 1\}.$$

Example 0.5.

- $\mathbb{Z}_2 = \{\overline{0}, \overline{1}\}$; $U_2 = \{\overline{1}\}$; $\phi(2) = 1$.
- $\mathbb{Z}_3 = \{\overline{0}, \overline{1}\,\overline{2}\}$; $U_3 = \{\overline{1}, \overline{2}\}$; $\phi(3) = 2$.
- $\mathbb{Z}_4 = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}\}$; $U_4 = \{\overline{1}, \overline{3}\}$; $\phi(4) = 2$.
- $\mathbb{Z}_5 = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}\}$; $U_5 = \{\overline{1}, \overline{2}, \overline{3}, \overline{4}\}$; $\phi(5) = 4$.
- $\mathbb{Z}_6 = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}, \overline{5}\}$; $U_6 = \{\overline{1}, \overline{5}\}$; $\phi(6) = 2$.
- p prime; $\mathbb{Z}_p = \{\overline{0}, \overline{1}, \dots, \overline{(p-1)}\}$; $U_p = \{\overline{1}, \dots, \overline{(p-1)}\}$; $\phi(p) = p - 1$.

As background for Fermat's little theorem and Euler's theorem, we want to talk about a few elementary properties of groups. We take a leisurely stroll.

Let G be a group and $g \in G$. The axioms for a group guarantee the existence of an inverse, that is an element h so that $g * h = h * g = e$. While the axioms don't say, the inverse is unique, for if k is another inverse (i.e., $g * k = k * g = e$), then

$$h = h * e = h * (g * k) = (h * g) * k = e * k = k,$$

showing they must be equal.

Denote the unique inverse of g by g^{-1}; we now extend this notation. For $k > 0$, denote by g^k the element $\underbrace{g * g * \cdots * g}_{k \text{ times}}$, $g^0 = e$, the identity, and for $k < 0$, let g^k be the inverse of $g^{|k|}$. With this shorthand convention, g^k satisfies the usual rules for exponents: $g^k g^\ell = g^{k+\ell}$ and $(g^k)^\ell = g^{k\ell}$.

Now suppose that G is a finite group, and consider the elements

$$g, g^2, g^3, \ldots, g^{r+1}$$

where $r = |G|$ is the order (cardinality) of the group G. Since there are $r+1$ elements in the list all of which are in G and only r distinct elements in G, two elements in the list must be the same, that is

$$g^i = g^j \quad \text{for some } 1 \leq i < j \leq r+1.$$

Multiplying both sides by g^{-1}, we see that $g^{j-i} = g^i g^{-i} = e$, and note that $1 \leq j - i \leq r$. We summarize this as a little proposition.

Proposition. *Let G be a finite group. Then for any element $g \in G$ there is a positive integer $n \leq |G|$ with $g^n = e$.*

Definition. We define the order of an element g, denoted $|g|$, as the smallest positive integer m so that $g^m = e$, if one exists. If no such positive integer exists, we say g has infinite order.

For example consider the group $G = U_9$ which has order $\phi(9) = 6$: $U_9 = \{\overline{1}, \overline{2}, \overline{4}, \overline{5}, \overline{7}, \overline{8}\}$, with $e = \overline{1}$.

$$\begin{aligned}
&\overline{1}^1 = \overline{1}, \text{ so } |\overline{1}| = 1 \\
&\overline{2}^1 = \overline{2},\ \overline{2}^2 = \overline{4},\ \overline{2}^3 = \overline{8},\ \overline{2}^4 = \overline{7},\ \overline{2}^5 = \overline{5},\ \overline{2}^6 = \overline{1} \text{ so } |\overline{2}| = 6 \\
&\overline{4}^1 = \overline{4},\ \overline{4}^2 = \overline{7},\ \overline{2}^3 = \overline{1} \text{ so } |\overline{4}| = 3 \\
&\overline{5}^1 = \overline{5},\ \overline{5}^2 = \overline{7},\ \overline{5}^3 = \overline{8},\ \overline{5}^4 = \overline{4},\ \overline{5}^5 = \overline{2},\ \overline{5}^6 = \overline{1} \text{ so } |\overline{5}| = 6 \\
&\overline{7}^1 = \overline{7},\ \overline{7}^2 = \overline{4},\ \overline{7}^3 = \overline{1} \text{ so } |\overline{7}| = 3 \\
&\overline{8}^1 = \overline{8},\ \overline{8}^2 = \overline{1}, \text{ so } |\overline{8}| = 2
\end{aligned}$$

An important theorem in group theory which we will prove if time permits is

Theorem. *Let G be a finite group, and $g \in G$. Then $g^{|G|} = e$.*

That is an element raised to the order of the group is the identity. We see this in our example above since the orders of all the elements divide 6, the order of U_9. In fact this is a fact we can prove.

Proposition. *Let G be a group and suppose that for some positive integer m, $g^m = e$. Then $|g| \mid m$. In particular in a finite group, we have that $|g| \mid |G|$.*

Proof. Let $d = |g|$. Then $g^d = g^m = e$ and $d4$ is by definition the smallest positive integer with that property. Use the division algorithm and write $m = dq + r$ with $0 \leq r < d$. Observe that $e = g^m = (g^d)^q g^r = g^r$, so if $0 < r < d$, it would violate that d is the order of g, thus $r = 0$ and we have $d \mid m$.

The second statement now follows since, by our unproven theorem, $g^{|G|} = e$. □

Indeed the theorem above gives one statement of Euler's theorem:

Theorem (Euler). *Let $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. Then $a^{\phi(n)} \equiv 1 \pmod{n}$.*

Note that Euler's theorem is simply the statement that if $\overline{a} \in U_n$, then $\overline{a}^{|U_n|} = \overline{1}$, that is a special case of the general result for finite groups. Since we have not proven the general result, we given a separate proof of Euler's theorem, but make the point that this is a good example of the power of abstraction.

Indeed Euler's theorem is itself a generalization of Fermat's little theorem:

Theorem (Fermat's little theorem). *Let p be a prime and a an integer with $p \nmid a$, then $a^{p-1} \equiv 1 \pmod{p}$. Equivalently, for any integer a, we have $a^p \equiv a \pmod{p}$.*

To prove Euler's theorem, we need to produce two different, but related sets of representatives for U_n. We start with

Proposition. *Suppose that $U_n = \{\overline{b}_1, \ldots, \overline{b}_{\phi(n)}\}$, and that a is an integer with $\gcd(a, n) = 1$. Then $U_n = \{\overline{ab}_1, \ldots, \overline{ab}_{\phi(n)}\}$. Equivalently, for each i with $1 \leq i \leq \phi(n)$, there is a unique j with $1 \leq j \leq \phi(n)$ so that $ab_i \equiv b_j \pmod{n}$.*

Before proving this, we give an example:

Example 0.6. $U_{10} = \{\overline{1}, \overline{3}, \overline{7}, \overline{9}\}$. Let $a = 13$ which is clearly coprime to 10. The proposition says that $U_{10} = \{\overline{13}, \overline{39}, \overline{91}, \overline{117}\} = \{\overline{3}, \overline{9}, \overline{1}, \overline{7}\}$. So we get the same classes, just in a different order.

Proof. First observe that each class $\overline{ab}_i \in U_n$. We could *reprove* this with congruences, but we have already established that U_n is a group under multiplication, and so closed under multiplication. Since $\gcd(a, n) = 1$, we have $\overline{a} \in U_n$, so $\overline{a}\overline{b}_i = \overline{ab}_i \in U_n$ by closure.

Next we observe that all the elements are distinct. Again, I do this using group theory:

Suppose that $\overline{ab_i} = \overline{ab}_j$, then $\overline{a}\overline{b}_i = \overline{a}\overline{b}_j$. Multiplying both sides by the inverse of $\overline{a}$ (U_n is a group!), we see that $\overline{b}_i = \overline{b}_j$, so all the representatives are distinct (i.e. different congruence classes), and since there are the correct number, they fill out all of U_n. □

Theorem (Euler). *Let $a, n \in \mathbb{Z}$ with $n \geq 1$ and $\gcd(a, n) = 1$. Then $a^{\phi(n)} \equiv 1 \pmod{n}$.*

Proof. Write $U_n = \{\overline{b}_1, \ldots, \overline{b}_{\phi(n)}\}$. Then as in the proposition, $U_n = \{\overline{ab}_1, \ldots, \overline{ab}_{\phi(n)}\}$. Since these are the same elements, in possibly a different order, their products are equal, that is

$$\overline{b}_1\overline{b}_2 \cdots \overline{b}_{\phi(n)} = \overline{ab}_1\overline{ab}_2 \cdots \overline{ab}_{\phi(n)} = \overline{a}^{\phi(n)}\overline{b}_1\overline{b}_2 \cdots \overline{b}_{\phi(n)}.$$

Since U_n is a group, the element $\overline{b}_1\overline{b}_2 \cdots \overline{b}_{\phi(n)}$ has an inverse in U_n. Multiplying both sides by the inverse, produces $\overline{a}^{\phi(n)} = \overline{1}$, or equivalently, $a^{\phi(n)} \equiv 1 \pmod{n}$. □

As an amusing corollary, we know that the congruence $ax \equiv b \pmod{n}$ has a unique solution mod n iff $\gcd(a, n) = 1$. Well, if it's unique, what is it? Since $\gcd(a, n) = 1$ we know $a^{\phi(n)} \equiv 1 \pmod{n}$ so multiplying both sides of the congruence by $a^{\phi(n)-1}$ yields $x \equiv a^{\phi(n)-1}b \pmod{n}$. Of course computing this solution may take time if $\phi(n)$ is large, which leads us to consider fast modular exponentiation.

Example 0.7. Let's compute $5^{123} \pmod{13}$. We consider Euler's (or Fermat's theorem) and realize that $5^{12} \equiv 1 \pmod{13}$, so $5^{123} = (5^{12})^{10}5^3 \equiv 5^3 \equiv 8 \pmod{13}$.

Example 0.8. Next observe that $341 = 11 \cdot 31$. We want to show that $2^{341} \equiv 2 \pmod{341}$. As we discussed, comparing this to Fermat's little theorem, we might wonder whether 341 is prime since it seems to behave like one when exponentiating with respect to the base 2. Indeed 341 is not prime, but this behavior earns it the name *pseudoprime to the base 2*.

Anyway since we know the factorization, we observe that $2^{341} \equiv 2 \pmod{341}$ iff $2^{341} \equiv 2 \pmod{11}$ and $2^{341} \equiv 2 \pmod{31}$. One direction is obvious and always true: If $a \equiv b \pmod{n}$ then $a \equiv \pmod{d}$ for any $d \mid n$. But the converse depends on the moduli 11 and 31 being coprime. We see that $11 \mid (a - b)$ and $31 \mid (a - b)$ says that $a - b = 11k$ for some integer. Since $31 \mid (a - b)$, we see $31 \mid 11k$ and since $\gcd(11, 31) = 1$, we have that $31 \mid k$, so that $k = 31k'$. Thus $a - b = 11k = 11 \cdot 31k'$ or $a \equiv b \pmod{3}41$.

By Euler's (or Fermat's) theorem, we know that $2^{10} \equiv 1 \pmod{11}$ and $2^{30} \equiv 1 \pmod{31}$. Thus

$$2^{341} \equiv (2^{10})^{34} \cdot 2^1 \equiv 2 \pmod{1}1,$$

and

$$2^{341} \equiv (2^{30})^{11} \cdot 2^{11} \equiv 2^{11} \pmod{31}.$$

To finish, we notice serendipitously that $2^5 = 32 \equiv 1 \pmod{31}$ so that $2^{11} \equiv (2^5)^2 \cdot 2^1 \equiv 2 \pmod{31}$.

Example 0.9. We do the last example again, but this time assuming we know nothing about the factorization. Being able quickly to perform modular exponentiation is critical to the implementation of the RSA cryptosystem.

Let's discuss this in general. We want to find $a^{341} \pmod{n}$ for some a and n. Their values are not really too important to the general plan. The key is to write the exponent in its base 2 expansion: $341 = 2^8 + 2^6 + 2^4 + 2^2 + 2^0$, so that

$$a^{341} \equiv a^{2^8+2^6+2^4+2^2+2^0} \equiv a^{2^8}a^{2^6}a^{2^4}a^{2^2}a^{2^0} \pmod{n}.$$

Now we observe that the terms a^{2^k} can be obtained by successive squaring, that is

$$a^{2^{k+1}} = a^{2^k \cdot 2} = (a^{2^k})^2.$$

So we compute:

$$\begin{aligned}
2^{2^0} &\equiv 2 \pmod{341}\\
2^{2^1} &\equiv (2^1)^2 \equiv 4 \pmod{341}\\
2^{2^2} &\equiv (2^{2^1})^2 \equiv 16 \pmod{341}\\
2^{2^3} &\equiv (2^{2^2})^2 \equiv 256 \pmod{341}\\
2^{2^4} &\equiv (2^{2^3})^2 \equiv 256^2 \equiv 64 \pmod{341}\\
2^{2^5} &\equiv (2^{2^4})^2 \equiv 64^2 \equiv 4 \pmod{341}\\
2^{2^6} &\equiv (2^{2^5})^2 \equiv 4^2 \equiv 16 \pmod{341}\\
2^{2^7} &\equiv (2^{2^6})^2 \equiv 16^2 \equiv 256 \pmod{341}\\
2^{2^8} &\equiv (2^{2^7})^2 \equiv 256^2 \equiv 64 \pmod{341}
\end{aligned}$$

Thus

$$2^341 \equiv 2^{2^8}2^{2^6}2^{2^4}2^{2^2}2^{2^0} \equiv 64 \cdot 16 \cdot 64 \cdot 16 \cdot 2 \equiv 2 \pmod{341}.$$

CHAPTER 5

Public Key Cryptography and RSA

To set a bit of notation, we let P denote a plaintext message, C denote the encrypted ciphertext, E an encryption algorithm, and D the associated decryption algorithm.

Any standard cryptographic system embodies the following properties: $P = D(E(P))$ and ideally both $E(P)$ and $D(E(P))$ are very fast to compute. In a public key cryptosystem (PKCS) we shall also demand that $E(D(C)) = C$ and that publicly revealing E does not reveal an easy was to deduce D. We shall see that the first of these additional conditions is what allows for electronic signatures, while the second is what permits secure communication between parties who have never met.

The security of every PKCS depends upon a task which is easy to do given privileged information (a trap door), but difficult to do otherwise. For RSA, the easy task is multiplying two integers together; the hard task is factoring the product. Slightly more to the point, RSA has as its basis the following idea. Take two large primes $p \neq q$ and form $n = pq$. What is $\phi(n)$? Well if we know p, q, the answer is trivial $\phi(n) = (p-1)(q-1)$, while if we do not know how to factor n, this is very hard. Indeed we shall show that knowing n and $\phi(n)$ is equivalent to knowing how to factor n.

Back to the task at hand, RSA is very easy to describe. As above, choose p, q, form $n = pq$, and $\phi(n)$. Choose (at random) a positive integer e with $\gcd(e, \phi(n)) = 1$. Random is fine as Euclid's algorithm is quick to use. On the hand, any prime not dividing $\phi(n) = (p-1)(q-1)$ will do. Many people like $65537 = 2^{2^4} + 1$, the last known Fermat prime (if it doesn't divide $p-1$ or $q-1$, and if does, go get another p, and q).

Once you have e, we know from general properties of linear congruences that there is a unique $d \pmod{\phi(n)}$ so that $ed \equiv 1 \pmod{\phi(n)}$; of course Euclid's algorithm finds this value quickly. To implement RSA, we convert our plaintext message into a numerical equivalent blocks $P < n$. Encryption is achieved by $C = E(P) \equiv P^e \pmod{n}$; of course fast modular exponentiation is needed here. Decryption is achieved by $P = D(C) \equiv C^d \pmod{n}$.

We need only verify that for $1 \leq P < n$, $P^{ed} \equiv P \pmod{n}$; this will tell both that $D(E(P)) = p$ and $C = E(D(C))$. The key is that $ed \equiv 1 \pmod{\phi)(n}$, which says that $ed = 1 + k\phi(n)$ for some integer k. Note that since n is the product of two primes, the only possibility for $\gcd(P, n)$ is $1 p, q, pq$, where the last is precluded since we assume $P < n$. The case of p or q are the same, so we have two cases:

Case 1: $\gcd(P, n) = 1$.

By Euler's theorem, $P^{\phi(n)} \equiv 1 \pmod{n}$, so

$$P^{ed} \equiv P^{1+k\phi(n)} = PP^{\phi(n)k} \equiv P \pmod{n}.$$

Case 1: $\gcd(P, n) > 1$; wlog $\gcd(P, n) = p$. Since p and q are relatively prime, $P^{ed} \equiv P \pmod{n}$ if and only if $P^{ed} \equiv P \pmod{p}$ and $P^{ed} \equiv P \pmod{q}$. Since $p \mid P$, we have $P^{ed} \equiv 0 \equiv P \pmod{p}$. Now $\gcd(P, n) = p$ means that $\gcd(P, q) = 1$. Thus we have:

$$P^{ed} \equiv P^1 P^{(q-1)(p-1)k} \equiv P\big(P^{(q-1)}\big)^{k(p-1)} \equiv P \mod q.$$

since $P^{(q-1)} \equiv 1 \pmod{q}$ from Fermat's little theorem (Euler's theorem for primes).

This shows that RSA will function correctly. We know look at the security of RSA. To break RSA, you need to compute d or extract eth roots modulo n. If your exponent e (which is public) is very small and n large compared to your plaintext P, one could simply trying taking eth roots in the real numbers to attempt to recover the plaintext; otherwise this is a difficult problem. To deduce d, you need to know the value of $\phi(n)$. We now show that knowledge of $\phi(n)$ yields knowledge of p and q.

We begin with the simple observation that

$$\phi(n) = (p-1)(q-1) = pq - (p+q) + 1 = n - (p+q) + 1.$$

so that knowing n and $\phi(n)$ tells us the value of $n - \phi(n) + 1 = p + q$. Also, (assuming without loss $p > q$)

$$p - q = \sqrt{(p-q)^2} = \sqrt{(p+q)^2 - 4pq} = \sqrt{(p+q)^2 - 4n}$$

is known. Given both $p+q$ and $p-q$, we immediately deduce p and q, so knowledge of $\phi(n)$ is equivalent to factoring $n = pq$.

In what follows we present successively more honest representations of how any PKCS (in particular RSA) is used. We start with overly simplified models to illustrate the essential features of a PKCS.

As in all literature concerning secure communication, our two players are always Alice and Bob. Each generates a pair of encryption and decryption algorithms, E_A, D_A, E_B, D_B. Each "publishes" their encryption algorithms in a public repository. Alice wants to send an encrypted message to Bob. She generates her plaintext message M_A, encrypts it with Bob's public encryption algorithm, and sends $E_B(M_A)$ to Bob over whatever insecure channel she likes. Bob receives, and extracts $M_A = D_B(E_B(M_A))$. So this is just like a secret key cryptosystem, except for the new ability to send to anyone with a published encryption key. No prior contact is required. For example, this is an essential feature whenever you want to order something online.

The issue of authentication is another critical issue in age when banking and legal transactions must take place over the internet. It is here that we see how easily a PKCS facilitates the signing of electronic documents. Once again we first take a simplistic approach. Alice wishes to send a message to Bob, which Bob can prove to a third party is indeed from Alice and also that the message has not been altered. Alice generates her message M_A, and signs it by creating $S_A = D_A(M_A)$, that is she uses her private decryption key and applies it to the plaintext. She then sends the message to Bob, as $E_B(S_A)$. Bob received the message and retrieves $S_A = D_B(E_B(S_A))$. Now Bob uses Alice's public encryption algorithm E_A to recover $M_A = E_A(S_A) = E_A(D_A(M_A))$. Bob can now hand M_A and S_A to a judge who verifies the message is from Alice as only Alice's public E_A are undo the signature. The judge is also confident the message M_A has not been tampered with, since to substitute a new message M'_A, for M_a, the forger would have to produce $D_A(M'_A)$ which he cannot.

A more realistic version would be for Alice to take her message and create a hash of it,$H_A = hash(M_A)$, which would serve as a digital fingerprint. She would form $S_A = D_A(H_A)$ sending the message M_A and signature S_A to Bob, would could verify its authenticity by hashing M_A and comparing to $E_A(S_A) = E_A(D_A(S_A)) = H_A$. Recall that hash functions take input of arbitrary length and produce output of fixed length (128 bit for MD5, 160 bit for SHA-1). There is an ongoing contest via NIST for a choice of SHA-2. Any hash function should be fast to compute, a one-way functions and be as collision-free as possible (MD5 fell victim to an attack to produce collisions), and generally changing one bit of data in the input stream should dramatically change the hash value. The analog of the birthday problem arises in considering how to defeat the use of hash functions.

Generally the modern use of a PKCS is to perform a secure key exchange, to provide a private session key to be used in a secret key cryptosystem for a electronic transaction. For example, you computer has on it a public and private key for a PKCS. When you go to Amazon's web site, your computer offers up your public key E. Amazon generates a random secret key k for an encrypted session, and sends $E(k)$ to your compute which obtains $k = D(E(k))$ and starts using the session key k in whatever encryption scheme Amazon wants.

CHAPTER 6

A little more group theory

1. Cayley tables

We need to talk about when two groups are (for all intents an purposes) the same. The technical word is that the groups are *isomorphic*. Below are the Cayley tables we worked out in class:

U_{10}	$\overline{1}$	$\overline{3}$	$\overline{7}$	$\overline{9}$
$\overline{1}$	$\overline{1}$	$\overline{5}$	$\overline{7}$	$\overline{9}$
$\overline{3}$	$\overline{3}$	$\overline{9}$	$\overline{1}$	$\overline{7}$
$\overline{7}$	$\overline{7}$	$\overline{1}$	$\overline{9}$	$\overline{3}$
$\overline{9}$	$\overline{9}$	$\overline{7}$	$\overline{3}$	$\overline{1}$

U_5	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$
$\overline{2}$	$\overline{2}$	$\overline{4}$	$\overline{1}$	$\overline{3}$
$\overline{3}$	$\overline{3}$	$\overline{1}$	$\overline{4}$	$\overline{2}$
$\overline{4}$	$\overline{4}$	$\overline{3}$	$\overline{2}$	$\overline{1}$

One observes that in mapping from $U_{10} \to U_5$, if we map $\overline{1} \mapsto \overline{1}$, $\overline{3} \mapsto \overline{2}$, $\overline{7} \mapsto \overline{3}$ and $\overline{9} \mapsto \overline{4}$, that the Cayley tables match exactly. A function $\varphi : U_{10} \to U_5$ defined as above would satisfy $\varphi(ab) = \varphi(a)\varphi(b)$ for all $a, b \in U_{10}$. This is an example of a *homomorphism* between groups, and a bijective homomorphism is an isomorphism.

We also computed Cayley tables of other groups of order four and found:

U_8	$\overline{1}$	$\overline{3}$	$\overline{5}$	$\overline{7}$
$\overline{1}$	$\overline{1}$	$\overline{3}$	$\overline{5}$	$\overline{7}$
$\overline{3}$	$\overline{3}$	$\overline{1}$	$\overline{7}$	$\overline{5}$
$\overline{5}$	$\overline{5}$	$\overline{7}$	$\overline{1}$	$\overline{3}$
$\overline{7}$	$\overline{7}$	$\overline{5}$	$\overline{3}$	$\overline{1}$

U_{12}	$\overline{1}$	$\overline{5}$	$\overline{7}$	$\overline{11}$
$\overline{1}$	$\overline{1}$	$\overline{5}$	$\overline{7}$	$\overline{11}$
$\overline{5}$	$\overline{5}$	$\overline{1}$	$\overline{11}$	$\overline{7}$
$\overline{7}$	$\overline{7}$	$\overline{11}$	$\overline{1}$	$\overline{5}$
$\overline{11}$	$\overline{11}$	$\overline{7}$	$\overline{5}$	$\overline{1}$

One observes that in mapping from $U_8 \to U_{12}$, if we map $\overline{1} \mapsto \overline{1}$, $\overline{3} \mapsto \overline{5}$, $\overline{5} \mapsto \overline{7}$ and $\overline{7} \mapsto \overline{11}$, that the Cayley tables match exactly. A function $\varphi : U_8 \to U_{12}$ defined as above is an isomorphism.

Next consider the Cayley table for $\mathbb{Z}_4$.

$\mathbb{Z}_4$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{3}$	$\overline{0}$	$\overline{1}$
$\overline{3}$	$\overline{3}$	$\overline{0}$	$\overline{1}$	$\overline{2}$

It doesn't obviously match either of the previous examples, but if we interchange the third and fourth row and column we see it matches with the first two examples:

$\mathbb{Z}_4$	$\overline{0}$	$\overline{1}$	$\overline{3}$	$\overline{2}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{3}$	$\overline{2}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{0}$	$\overline{3}$
$\overline{3}$	$\overline{3}$	$\overline{0}$	$\overline{2}$	$\overline{1}$
$\overline{2}$	$\overline{2}$	$\overline{3}$	$\overline{1}$	$\overline{0}$

Finally consider the Cayley table for $\mathbb{Z}_2 \times \mathbb{Z}_2$ (without the bars over the numbers)

$\mathbb{Z}_2 \times \mathbb{Z}_2$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$	$(0,0)$	$(0,1)$	(1.0)	$(1,1)$
$(0,1)$	$(0,1)$	$(0,0)$	$(1,1)$	$(1,0)$
$(1,0)$	$(1,0)$	$(1,1)$	$(0,0)$	$(0,1)$
$(1,1)$	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$

We see it matches with U_8 and U_{12}. In terms of isomorphisms, we have $U_{10} \cong U_5 \cong \mathbb{Z}_4$, and $U_8 \cong U_{12} \cong \mathbb{Z}_2 \times \mathbb{Z}_2$. Indeed, up to isomorphism there are only two nonisomorphic groups of order 4.

A bit of terminology. Let G be a group and $g \in G$. We denote by

$$\langle g \rangle = \{g^m \mid m \in \mathbb{Z}\}.$$

We easily check that $\langle g\rangle$ is a group called the *cyclic subgroup generated by* g. A group is called *cyclic* if there exists a $g \in G$ with $G = \langle g\rangle$. Thus $U_{10} = \langle\overline{3}\rangle = \langle\overline{7}\rangle$, $U_5 = \langle\overline{2}\rangle = \langle\overline{3}\rangle$, and $\mathbb{Z}_4 = \langle\overline{1}\rangle$ are all cyclic; we note the operation in $\mathbb{Z}_4$ is additive, so g^m means $g + g + \cdots + g$ (m times).

The other groups are not cyclic, all elements having order 1 or 2, and so can't fill out all of G.

We state without proof and important result in group theorem, a theorem of Lagrange.

Theorem. *Let G be a finite group and $H \subseteq G$ a subgroup (a subset which is also a group). Then $|H| \,\big|\, |G|$. That is the order of a subgroup divides the order of the group.*

Using this we show that

Corollary. *Let G be a finite group, $g \in G$ and put $H = \langle g\rangle$. Then $H = \{e = g^0, g, g^2, \ldots, g^{d-1}\}$, where $d = |g|$. In particular, the order of an element divides the order of the group.*

Proof. Let m be any integer. Using the division algorithm, write $m = dq + r$ where $d = |g|$ and where $0 \leq r < d$. Thus $g^m = g^d q g^r = g^r$, so every element of $\langle g\rangle$ is in $\{e = g^0, g, g^2, \ldots, g^{d-1}\}$. It only remains to show that no two elements in the list $g^0, g, g^2, \ldots, g^{d-1}$ are equal. If that were true, d would not be the smallest positive exponent so that $g^d = e$, which would provide a contradiction. □

Corollary. *Let G be a finite group with prime order p. Then G is cyclic.*

Proof. Let $g \in G$ with $g \neq e$, and put $H = \langle g\rangle$. Then $|H| > 1$, and by Lagrange $|H| \mid |G| = p$. This implies $|H| = p$ which means H is all of G. □

Example 1.1. Note that $(\mathbb{Z}_n, +)$ is a cyclic group (under addition) for any $n \geq 1$, so there exist cyclic groups of every finite order.

On the other hand there is only one cyclic group of a given order, that is up to isomorphism.

Proposition. *Let $G_1 = \langle g_1\rangle$ and $G_2 = \langle g_2\rangle$ both be cyclic groups of order n. Then $G_1 \cong G_2$, that is they are congruent.*

Proof. $G_1 = \{e, g_1, g_1^2, \ldots, g_1^{n-1}\}$ and $G_2 = \{e, g_2, g_2^2, \ldots, g_2^{n-1}\}$. Sending $g_1^k \mapsto g_2^k$ shows that the Cayley tables match establishing the isomorphism. □

Corollary. *Let G be a cyclic group of order n. Then $G \cong \mathbb{Z}_n$.*

Example 1.2. Classifying all finite groups. One large project would be to list all the isomorphism classes of finite groups. We start the process by listing them within a given order n.

- $n = 1$: $G = \{e\}$, called the trivial group.
- $n = 2, 3$: Both numbers are prime, so the groups are cyclic and isomorphic to (respectively) $\mathbb{Z}_2$, and $\mathbb{Z}_3$.
- $n = 4$: We have seen $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are not isomorphic. It turns out these are the only two, that is every group of order four is isomorphic to one of these.
- $n = 5$: Prime order, so cyclic, $G \cong \mathbb{Z}_5$.
- $n = 6$: First interesting case. Clearly $\mathbb{Z}_6$ is possible, and it turns out $\mathbb{Z}_2 \times \mathbb{Z}_3 \cong \mathbb{Z}_6$. On the other hand for homework you should $S_3 \cong D_3$ are groups of order six, but not abelian so certainly not isomorphic to $\mathbb{Z}_6$.
- $n = 7$: prime again $\mathbb{Z}_7$.
- $n = 8$ Five groups: $\mathbb{Z}_8, \mathbb{Z}_4 \times \mathbb{Z}_2, \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ are the abelian ones. D_4 the symmetries of the square is another. The last is called the quaternion group.

2. Fundamental Theorem of Finite Abelian Groups

The structure of finite abelian groups is particularly easy to describe, though we take a somewhat iterative approach to the final result. As preliminary, we talk about what appears an unrelated topic: the partitions of a positive integer n.

We say that $\{n_1, \ldots, n_k\}$ is a partition of the positive integer n if $n = n_1 + \cdots + n_k$ and $n_1 \geq n_2 \geq \cdots \geq n_k > 0$, and we let $p(n)$ be the number of partitions of n. Thus $p(5) = 7$ since there are 7 partitions of the number 5:

$$5 = 5; \quad 4 + 1; \quad 3 + 2; \quad 3 + 1 + 1 \quad ; 2 + 2 + 1; \quad 2 + 1 + 1 + 1; \quad 1 + 1 + 1 + 1 + 1.$$

Our first theorem is for abelian groups of prime-power order.

Theorem. *Let q be a prime, and $n \geq 1$. Then up to isomorphism, there are $p(n)$ abelian groups of order q^n.*

Said another way, there are only $p(n)$ distinct Cayley tables you can write down for an abelian group of order q^n. What is striking about this result is that it depends only on the exponent n and not on the prime q. Moreover, the proof shows not only that the partition function $p(n)$ counts the number of abelian groups, but that the partitions of n tell you their shape.

EXAMPLE 2.1. We characterize all the abelian groups of order q^n in terms of partitions as follows:

$$\begin{aligned}
5 &\longleftrightarrow \mathbb{Z}_{q^5}\\
4+1 &\longleftrightarrow \mathbb{Z}_{q^4}\times\mathbb{Z}_q\\
3+2 &\longleftrightarrow \mathbb{Z}_{q^3}\times\mathbb{Z}_{q^2}\\
3+1+1 &\longleftrightarrow \mathbb{Z}_{q^3}\times\mathbb{Z}_q\times\mathbb{Z}_q\\
2+2+1 &\longleftrightarrow \mathbb{Z}_{q^2}\times\mathbb{Z}_{q^2}\times\mathbb{Z}_q\\
2+1+1+1 &\longleftrightarrow \mathbb{Z}_{q^2}\times\mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q\\
1+1+1+1+1 &\longleftrightarrow \mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q
\end{aligned}$$

THEOREM. *Let $N \geq 2$ be an integer with prime factorization $N = q_1^{e_1}\cdots q_r^{e_r}$. Then every abelian group G of order N is isomorphic to a direct product $G \cong G(q_1)\times\cdots\times G(q_r)$ where $G(q_i)$ is an abelian group of order $q_i^{e_i}$, described in the previous theorem. Thus up to isomorphism, there are $p(e_1)p(e_2)\cdots p(e_r)$ abelian groups of order N.*

It is useful in analyzing abelian groups to note that

THEOREM. $\mathbb{Z}_m\times\mathbb{Z}_n\cong\mathbb{Z}_{mn}$ *if and only if* $\gcd(m,n)=1$.

EXAMPLE 2.2. Classify all abelian groups of order p^2q^3 where p and q are distinct primes.

There are $p(2) = 2$ abelian groups of order p^2: $\mathbb{Z}_{p^2}$ and $\mathbb{Z}_p\times\mathbb{Z}_p$. There are $p(3) = 3$ abelian groups of order q^3: Z_{q^3}, $\mathbb{Z}_{q^2}\times\mathbb{Z}_q$, and $\mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q$. Thus there are $p(2)p(3) = 6$ abelian groups of order p^2q^3, and they are (the left-hand column being the familiar one):

$$\begin{aligned}
\mathbb{Z}_{p^2}\times\mathbb{Z}_{q^3} &\cong \mathbb{Z}_{p^2q^3}\\
\mathbb{Z}_{p^2}\times\mathbb{Z}_{q^2}\times\mathbb{Z}_q &\cong \mathbb{Z}_q\times\mathbb{Z}_{p^2q^2}\\
\mathbb{Z}_{p^2}\times\mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q &\cong \mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_{p^2q}\\
\mathbb{Z}_p\times\mathbb{Z}_p\times\mathbb{Z}_{q^3} &\cong \mathbb{Z}_p\times\mathbb{Z}_{pq^3}\\
\mathbb{Z}_p\times\mathbb{Z}_p\times\mathbb{Z}_{q^2}\times\mathbb{Z}_q &\cong \mathbb{Z}_{pq}\times\mathbb{Z}_{pq^2}\\
\mathbb{Z}_p\times\mathbb{Z}_p\times\mathbb{Z}_q\times\mathbb{Z}_q\times\mathbb{Z}_q &\cong \mathbb{Z}_p\times\mathbb{Z}_{pq}\times\mathbb{Z}_{pq}
\end{aligned}$$

3. Rational Points on curves in affine and projective space

This section steals heavily from [**6**].

We have seen the definition of the projective line in our homework, but without much geometric motivation. Here we shall discuss the projective plane with at least one motivation being geometric: can we design a plane in which any two lines (parallel or not) intersect in a unique point and which contains the usual cases of intersecting lines as a subset?

We also take a slightly algebraic approach which motivates the structure of projective space as homogeneous coordinates.

Previously we have exploited a correspondence between rational points on the unit circle $x^2 + y^2 = 1$ with Pythagorean triples. We extend that correspondence to the Fermat curves $x^n + y^n = 1$ for $n \geq 2$. Let $(a/b, c/d)$ be a rational point on the Fermat curve with $\gcd(a, b) = \gcd(c, d) = 1$ and without loss of generality $b, d > 0$.

Then $(a/b)^n + (c/d)^n = 1$ implies that $(ad)^n + (bc)^n = (bd)^n$. Thus

$$b^n \mid (ad)^n \text{ and } \gcd(a, b) = 1 \implies b^n \mid d^n \implies b \mid d$$
$$d^n \mid (bc)^n \text{ and } \gcd(c, d)) = 1 \implies d^n \mid b^n \implies d \mid b.$$

Thus $b = \pm d$, but since both are positive, we have $b = d$. So all rational points on $x^n + y^n = 1$ have the form $(a/c, b/c)$ with $\gcd(a, c) = \gcd(b, c) = 1$ and $c > 0$. Thus we have a correspondence in one direction:

$$\left(\frac{a}{c}, \frac{b}{c}\right) \text{ on } x^n + y^n = 1 \mapsto (a, b, c) \text{ with } a^n + b^n = c^n.$$

Conversely, an integer solution $a^n + b^n = c^n$ with $c \neq 0$ corresponds to a rational point $(a/c, b/c)$ on $x^n + y^n = 1$, but this correspondence is far from one-to-one. In particular, every point of the form (at, bt, ct), $t \neq 0$, corresponds to the same rational point $(a/c, b/c)$, so if we want to get a one-to-one correspondence we would be forced to identify all the points (at, bt, ct), with $t \neq 0$. This of course leads to the definition of projective space and homogeneous coordinates.

Before making that definition, there is another issue with which to deal and that is solutions to $a^n + b^n = c^n$ where $c = 0$. While we can easily dismiss the case when $a = b = c = 0$ as irrelevant, when n is odd there are nontrivial solutions, e.g., $a^n + (-a)^n = 0$ for any nonzero a. Under the conjectured correspondence these would seem to correspond to "rational" points $(a/0, -a/0) = (\infty, \infty)$, certainly not a typical point on the Fermat curve.

For $n \geq 1$ we define projective n-space over a field F (e.g., $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}_p$) as follows:

Let $S = \{(a_1, a_2, \ldots, a_{n+1}) \in F^{n+1} \mid (a_1, a_2, \ldots, a_{n+1}) \neq (0, \ldots, 0)\}$. Define a relation on S by $(a_1, a_2, \ldots, a_{n+1}) \sim (b_1, \ldots, b_{n+1})$ if and only if there is a nonzero scalar $\lambda \in F$ so that $(b_1, \ldots, b_{n+1}) = \lambda(a_1, \ldots, a_{n+1})$. As in the homework, we easily check that $\sim$ is an equivalence relation and we denote by $[a_1, \ldots, a_{n+1}]$ the equivalence class of $(a_1, \ldots, a_{n+1})$ in S. Projective n-space is this set of equivalence classes:

$$\mathbb{P}^n(F) = \{[a_, \ldots, a_{n+1}] \mid (a_1, \ldots, a_{n+1}) \in S\}.$$

In particular, the projective plane $\mathbb{P}^2(F)$ is simply the set of classes $[a, b, c]$ with $(a, b, c) \in F^3$, $(a, b, c) \neq (0, 0, 0)$. Geometrically, the points in the projective plane are in one-to-one correspondence with the lines through the origin in F^3. As in the case of the projective line, it is often convenient to think of a distinguished set of representatives of the points in the projective plane. For example, and point $[a, b, c]$ either satisfies $c = 0$ or not. If $c \neq 0$, then $[a, b, c] = [a/c, b/c, 1]$ affording a correspondence with "affine" points $(a/c, b/c) \in \mathbb{A}^2(F) := F^2$ as we saw in the Fermat curve example. On the other hand points where $c = 0$ do not correspond to points in the affine plane. They are points on the projective line $z = 0$, the so-called line at infinity.

Indeed a closer look at such points reveals a bit more structure: A point $[a, b, 0] \in \mathbb{P}^2(F)$ means that a and b cannot both be zero, and

$$[a, b, 0] = \begin{cases} [a/b, 1, 0] & \text{if } b \neq 0 \\ [a, 0, 0] = [1, 0, 0] & \text{if } b = 0. \end{cases}$$

Thus we see that the set of point $[a, b, 0]$ is just a copy of the projective line $\mathbb{P}^1(F)$, so that $\mathbb{P}^2(F) = \mathbb{A}^2(F) \cup \mathbb{P}^1(F)$. This kind of decomposition generalizes easily to $\mathbb{P}^n(F)$ where we obtain a "hyperplane" at infinity as the extra set of points.

Now lets get a sense of affine and projective curves.

We all know what it means for an affine point (a, b) to lie on an affine curve $y^2 = x^3 + \alpha x^2 + \beta x + \gamma$, namely that $b^2 = a^3 + \alpha a^2 + \beta a + \gamma$, but projective points are equivalence classes and so whatever equation the coordinates of the point $[a, b, c]$ satisfy must also be satisfied by $[at, bt, ct]$ for all nonzero t. This means we need to "homogenize" the equation, introducing a new variable so that each summand becomes a polynomial of the same degree, thus

$$y^2 = x^3 + \alpha x^2 + \beta x + \gamma \mapsto y^2 z = x^3 + \alpha x^2 z + \beta x z^2 + \gamma z^3.$$

A significant observation is that substituting $z = 1$ into the equation of the projective curve produces the equation of the original affine curve.

Let's look at a few examples to make these ideas clearer.

Consider the points of intersection of the affine curves $x = y^2$ and $y = -3$. As this is the intersection of a line and a conic, we expect at most two points, and indeed there is only one affine point $(9, -3)$.

We homogenize (projectivize) the curves obtaining $xz = y^2$ and $y = -3z$. Equating the two gives $xz = 9z^2$ or $z(9z - x) = 0$. So either $z = 0$ or $x = 9z$ or both. First we conclude no solutions result from both conditions being true. $z = 0$ implies $x = 9z = 0$ and $y = -3z = 0$, and $[0, 0, 0]$ is not a point in projective space. So we have two cases $z \neq 0$ and $x = 9z$, or $z = 0$. So $x = 9z, y = -3z$ and $z \neq 0$ gives us the single projective point $[9z, -3z, z] = [9, -3, 1]$ which corresponds to our affine solution $(9, -3)$. That leave the case where $z = 0$. In that case our we have $y^2 = xz = 0 = -3z$, so we have $y = z = 0$ and x is arbitrary (but not zero), so we gain one more point $[x, 0, 0] = [1, 0, 0]$ of intersection which lies on the line at infinity.

Let's look at the intersection of the line parallel lines $y = 3x$ and $y = 3x + 1$. Of course there are no points of intersection in the affine plane, so we look projectively. The corresponding projective lines are $y = 3x$ and $y = 3x + z$. Equating, we see $3x = 3x + z$, so

$z = 0$ (which is good since it says the only possible solutions are on the line at infinity since we know there are no affine solutions). So we have $z = 0$ and $y = 3x$, which gives the single point $[x, 3x, 0] = [1, 3, 0]$ as the point of intersection of these projective lines.

As a last example we look at the intersection of the cubic $y = x^3$ and the line $y = x + 6$. We would like to see three points of intersection, but where are they?

We see immediately that $x^3 = x + 6$ iff $x^3 - x - 6 = (x - 2)(x^2 + 2x + 3) = 0$, and the quadratic has no real roots, though it has the complex roots $-1 \pm i\sqrt{2}$. At any rate the point $(2, 8)$ is a point on the affine curve so $[2, 8, 1]$ should be a point on the projective curves: $yz^2 = x^3$ and $y = x + 6z$. Looking at the line at infinity $z = 0$, we see $x = 0$ and hence $y = 0$, so there are no projective points on two curves which were not affine. Indeed we see that the solutions are $(2, 8)$ if we look in $\mathbb{A}^2(\mathbb{R})$, though we get three points $(2, 8), (1 \pm i\sqrt{2}, 7 \pm i\sqrt{2})$ if we look in $\mathbb{A}^2(\mathbb{C})$.

So the moral of the story is that we have any hope of finding an environment in which a curve of degree m and a curve of degree n intersect in mn points we need to look in $\mathbb{P}^2(\mathbb{C})$, that is the projective plane over the algebraically closed field $\mathbb{C}$.

Another nice perspective given in [**6**] is to ask just how many points would one need to add to $\mathbb{A}^2(\mathbb{R})$ to ensure that any two lines intersect. Their argument is as follows (sans pictures). Let $L_1 \neq L_2$ be two parallel lines and let P be the point "at infinity" which we will add to make them intersect. Analogously, let $L_1' \neq L_2'$ be two other parallel lines, and let P' be the point at infinity at which they will intersect. Now if L_1 and L_1' are not parallel, then they intersect in the affine plane, say at a point Q. If $P = P'$, then $L_q \cap L_1'$ contains both the point Q and $P = P'$ giving two distinct points of intersection, which is impossible. So for each direction of line in the plane we need a distinct point at infinity. A line in the affine plane either has slope $m \in \mathbb{R}$ or is a vertical line, and we add the points $[1, m, 0]$ $(m \in \mathbb{R})$ and the point $[0, 1, 0]$ to the affine line to make our projective plane.

Now as an exercise we look at two affine lines $a_i x + b_i y + c_i = 0$ in $\mathbb{R}^2$. These lines are parallel iff $a_1 b_2 - a_2 b_1 = 0$ since they are either both vertical $(b_i = 0)$ or not and hence have equal slopes: $-a_i/b_i$. Consider the corresponding projective lines: $a_i x + b_i y + c_i z = 0$. If they are not parallel, multiplying the first equation by b_2 and the second by b_1 results in

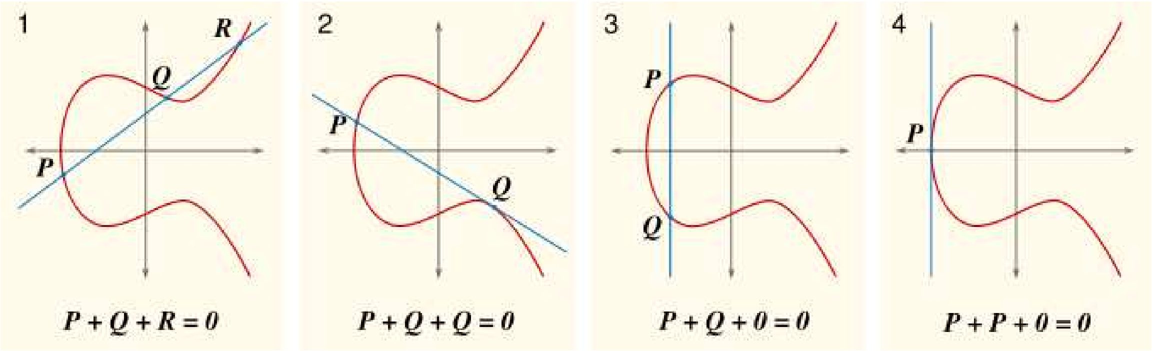
$$a_1 b_2 x + b_1 b_2 y + b_2 c_1 z = 0$$
$$a_2 b_1 x + b_1 b_2 y + c_2 b_1 z = 0$$

Note at least one of the b_i are nonzero (else both lines vertical, hence parallel). Subtracting, we see that $(a_1 b_2 - a_2 b_1)x + (b_2 c_1 - b_1 c_2)z = 0$. Since the lines are not parallel, we know the coefficient of x is not zero, so we may solve for x in terms of z which then will give y in terms of z, and we obtain one point of intersection which takes place (as usual) in the affine plane (set $z = 1$).

If the lines are parallel (but not equal) they have the form $ax+by+c=0$ and $ax+by+c'=0$ with $c \neq c'$. The projective lines have the form $ax+by+cz=0$ and $ax+by+c'z=0$. Subtracting, we get $(c-c')z=0$ which implies that $z=0$ since $c \neq c'$. Thus we have $z=0$ and $ax+by=0$ which yields the unique point $[b,-a,0]$ in the projective plane.

4. The Group Law on an elliptic curve

There are a number of ways to define a group structure on the set of points on an elliptic curve with coordinates in a fixed field F, but the most intuitive to understand is via a geometric construction for which our brief foray into projective space has adequately equipped us. So to set our intuition we first work over the field of real numbers where we can draw pictures. To save me a bit of time, I have stolen the following images from `http://en.wikipedia.org/wiki/Elliptic_curve`.



The identity for the group will be the point at infinity $\mathcal{O}=[0,1,0]$; all other points P are affine and have the form $P=[x,y,1]$. For any point $[x,y.z]$ on the curve, we denote by its additive inverse the point $[x,-y,z]$, so $-\mathcal{O}=\mathcal{O}$ and otherwise $-P$ is simply the reflection across the x-axis of the point P. This is indicated in picture 3, with $P+Q=\mathcal{O}$. The rest of the law is described by pictures 1,2, 4. Given two points P and Q on the elliptic curve, we first draw the line between P and Q. In the case that $P=Q$, we draw the tangent line. Picture 1 shows the generic situation where the line through P and Q meets the curve in a distinct third point R. The law says that $P+Q+R=\mathcal{O}$, that is $P+Q=-R$ the point obtained by reflecting the point R across the x-axis. In picture 2, we see the tangent line at Q intersecting at the "third" point Q, so that $Q+Q=2Q=-P$. Picture 4 shows what happens when we double a point on the x-axis, namely $2P=\mathcal{O}$.

In our case we shall consider elliptic curves over fields F like $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}_p$ ($p \neq 2,3$ which require more effort). In these cases a curve can be assumed in the form $y^2=x^3+ax+b$ with $a,b,\in F$. The discriminant of this cubic is $D=4a^3-27b^2$ and the curve defines an elliptic curve when the cubic has distinct roots, and that is measured by the discriminant not being zero (think about the quadratic formula for comparison).

We have seen that the projective (homogenized) curve $y^2z=x^3+axz^2+bz^3$ has one point $\mathcal{O}=[0,1,0]$ (where all vertical lines intersect the curve), and all other points are affine.

Suppose we want to find the sum of two points P_1 and P_2 with $P_i = [x_i, y_i, 1]$. Let L be the line between the two points (the tangent line if $P_1 = P_2$). The case of vertical lines is really described by pictures 3,4, so we assume the line L is not vertical. The slope of the line is

$$m = \begin{cases} \frac{y_2 - y_1}{x_2 - x_1} & \text{if } x_1 \neq x_2 \\ \\ \frac{3x_1^2 + a}{2y_1} & \text{if } x_1 = x_2, \end{cases}$$

where the slope of the tangent line is obtained by differentiation implicitly the equation $y^2 = x^3 + ax + b$, and evaluating at (x_1, y_1). So the equation the line L through P_1 and P_2 is

$$y - y_1 = m(x - x_1).$$

Solving for y and substituting into $y^2 = x^3 + ax + b$, we obtain

$$(m(x - x_1) + y_1)^2 = x^3 + ax + b.$$

Expanding and collecting terms we find

$$x^3 - m^2 x^2 + (*)x + (*) = 0,$$

with the (*)s not needed for what we do below.

Now the roots of the cubic are precisely the x-coordinates of the three points of intersection, x_1, x_2 (possibly equal), and x_3 the x-coordinate of the third point of intersection of the line and the elliptic curve. Thus,

$$x^3 - m^2 x^2 + (*)x + (*) = (x - x_1)(x - x_2)(x - x_3) = x^3 - (x_1 + x_2 + x_3)x^2 + (*)x + (*),$$

from which we conclude (comparing coefficients of x^2) that $x_3 = m^2 - x_1 - x_2$. Thus $y_3 = m(x_3 - x_1) + y_1$. So the point of intersection of the line L with the curve is (x_3, y_3), so $P_1 + P_2 = [x_3, -y_3, 1]$ by our geometric rule.

Summary of Group Law: Let $\mathcal{O} = [0, 1, 0]$ and $P_i = [x_i, y_i, 1]$.

(1) $-\mathcal{O} = [0, -1, 0] = [0, 1, 0] = \mathcal{O}$.
(2) $-P_i = [x_i, -y_i, 1]$.
(3) $P_i + \mathcal{O} = \mathcal{O} + P_i = P_i$.
(4) $P_2 = -P_1$ iff $P_1 + P_2 = \mathcal{O}$,
(5) If $P_2 \neq -P_1$, then $P_1 + P_2 = P_3 = [x_3, y_3, 1]$, where

$$x_3 = m^2 - x_1 - x_2,\ -y_3 = m(x_3 - x_1) + y_1 \text{ and where } m = \begin{cases} \frac{y_2 - y_1}{x_2 - x_1} & \text{if } x_1 \neq x_2 \\ \\ \frac{3x_1^2 + a}{2y_1} & \text{if } x_1 = x_2. \end{cases}$$

These rules make the set of points on the curve into an abelian group. Note that if the coordinates of P_1 and P_2 all lie in a fixed field (e.g., $\mathbb{Q}$ or $\mathbb{Z}_p$), then the coordinates of $P_1 + P_2$ also lie in that same field. So we may write $E(F)$ for the group of F-rational points on the curve (the points with coordinates in F) and this set of points is also an abelian group.

We note that the formulas for the slope given above still make sense even when we are in the finite field $\mathbb{Z}_p$. Both formulas require us to divide, but we are in situations where the denominator is nonzero. This means the denominator is in U_p all of whose elements have multiplicative inverses. We demonstrate with an example.

EXAMPLE 4.1. Consider the example [**6**] of the cubic curve $y^2 = x^3 + x + 1$ over the field $\mathbb{Z}_5$. The discriminant is $4a^3 + 27b^2 = 4 + 27 = 31 \not\equiv 0 \pmod 5$, so this is an elliptic curve over $\mathbb{Z}_5$. We check directly that

$$E(\mathbb{Z}_5) = \{\mathcal{O}, (0, \pm 1), (2, \pm 1), (3, \pm 1), (4, \pm 1)\},$$

so $E(\mathbb{Z}_5)$ is an abelian group of order 9, which we know by the fundamental theorem is isomorphic to $\mathbb{Z}_3 \times \mathbb{Z}_3$ or to $\mathbb{Z}_9$.

Consider the point $P = (0, 1) = [0, 1, 1]$ on the curve. To find the order of P, we need to find the smallest positive integer k so that $kP = P + P + \cdots + P(k \text{ times}) = \mathcal{O}$. We begin by computing $2P$ using our formulas, though we know that by Lagrange' theorem the order must be 3 or 9.

To double the point P we use the slope of the tangent line in our formulas above:

$$m = \frac{3x_2^2 + a}{2y_1} = \frac{3 \cdot 0 + 1}{2 \cdot 1} = \frac{1}{2} = 3 \text{ in } \mathbb{Z}_5.$$

Thus, $2P = (x_3, y_3) = (4, 2)$ (by our formulas). Similarly, we compute:

$$\begin{aligned} P &= [0, 1, 1], \\ 2P &= [4, 2, 1], \\ 3P &= [2, 1, 1] \neq \mathcal{O}, \\ 4P &= [3, -1, 1], \\ 5P &= [3, 1, 1]. \end{aligned}$$

It is clear that $4P = -5P$, so of course $4P + 5P = \mathcal{O}$, and since we know the order of P must divide 9, and is not 3, P has order 9. This makes our group of points cyclic with P for a generator.

An important question is how large $E(\mathbb{Z}_p)$ can be. For simplicity, let's restrict to odd primes. As we plug in the values $x = 0, 1, 2, \ldots, p - 1$ into $y^2 = f(x)$ we have several possibilities. If $f(x_0)$ is not a square in $\mathbb{Z}_p$, the we have no points of the form (x_0, y) on the curve; if $f(x_0) = 0$, we get one, namely $(x_0, 0)$ and if $f(x_0) = y_0^2$ is a nonzero square, we get two: $(x_0, \pm y_0)$.

As we look at the multiplicative group U_p, which has order $p-1$, we see that $k^2 = (-k)^2$ for $k = 1, 2, \ldots, (p-1)/2$. These represent all possible squares in U_p, and they are all distinct, since $a^2 \equiv b^2 \pmod{p}$ iff $(a-b)(a+b) \equiv 0 \pmod{p}$. So $p \mid (a-b)(a+b)$ which means (since p is prime) that $p \mid (a-b)$ or $(a+b$, which is to say $a \equiv \pm b \pmod{p}$.

All this says is that among the elements of U_p, half of them are squares and half are nonsquares. Indeed the squares and nonsquares of U_p, or of $\mathbb{Z}_p$ are so important to number theory, they have special names. If $a \in U_p$, we say a is a *quadratic residue* if a is a square in U_p, and a quadratic residue, otherwise.

Returning to our question of the size of $E(\mathbb{Z}_p)$, we first offer a heuristic: Suppose that as x runs over all the values of $\mathbb{Z}_p$, the values of $f(x)$ are uniformly distributed modulo p.

So we expect, some x_0 for which $f(x_0) = 0$ (yielding 1 point on the curve), $(p-1)/2)$ of the points x in U_p yielding $f(x)$ a quadratic residue, contributing another $(p-1)$ points on the curve, plus the point at infinity for a total guess of $p+1$ points.

With that as a guess, based on a uniform distribution, we write $\#E(\mathbb{Z}_p) = p + 1 +$ (error term). That's all well and good, but is there any truth in this heuristic? That is answered by a theorem of Hasse which says the error is bounded in absolute value by $2\sqrt{p}$:

THEOREM. *(Hasse)*

$$-2\sqrt{p} \leq \#E(\mathbb{Z}_p) - p - 1 \leq 2\sqrt{p}.$$

In fact, a theorem of Deuring (1941) says that if we let $E_{a,b}$ denote the elliptic curve $y^2 = x^3 + ax + b$, then for any integer m with $p + 1 - 2\sqrt{p} < m < p + 1 + 2\sqrt{p}$, there exists $a, b \in \mathbb{Z}_p$ so that $\#E(\mathbb{Z}_p) = m$. A 1987 theorem of Henrik Lenstra says there are lots of them.

CHAPTER 7

Factoring Integers

While in theory, we know that every integer $n > 1$ can be factored uniquely into a product of primes, in practice this can be very hard to do. More to the point, we have said how the security of RSA rests upon the practical intractability of factoring large numbers. Of course, factoring is a recursive process, so we will always be content with starting with a given $n > 1$ and writing $n = ab$ with $1 < a, b < n$. If this is not possible, then n is a prime (also a good thing to know), but if it is, then one of the factors a or b must be $\leq \sqrt{n}$. So as first method of factoring we consider trial division.

1. Trial Division

As we have indicated above, a composite number n must have a divisor whose value is $\leq \sqrt{n}$, so given an integer $n > 1$, we can try dividing by $2, 3, 4, 5, 6 \ldots, , \lfloor\sqrt{n}\rfloor$. If none of these divides n, we know that n is prime, but this is not a particularly efficient method.

For example, if n is an integer with approximately 100 decimal digits and our computer can check 1 million trial divisions per second, it could take as long as 3.2×10^{37} years to check all the divisors. If we used a computer a million times faster, that is 1 trillion trial divisions per second, it could take up to 3.2×10^{31} years. Given that the estimated age of the universe is approximately 13.5 billion years (1.35×10^{10}) we are talking about a process whose length would be the cube of the age of the universe. Probably less than optimal.

In the next two sections, we consider two related factoring methods, the later of which is one method representing state of the art in this field. Of course, before one starts to try to factor a number, it is convenient to know it is actually composite. Fortunately the task of determining whether a number is prime is much simpler than that of factoring a composite (and a subject worth its own chapter or two), but times constrains us to assume that our given integer is composite (such as an RSA modulus), and go from there.

2. Pollard's $p - 1$ method

Pollard's $p - 1$ will be effective on those integers n which possess a prime divisor p, so that $p-1$ is itself the product of small primes. Of course if this is not the case, the algorithm will run a very long time and reveal nothing, which is why factoring is still as much an art as science. So we shall assume our n has such a prime divisor p.

The problem is, even if we make this assumption, how does this help? After all, we don't know p. Let's keep our eye on the fact that what we are after is a non-trivial divisor of n. One silly way to do that is to choose integers b at random and compute $\gcd(b,n)$. Typically we will get a gcd of 1 or n, but if we get a gcd d with $1 < d < n$, we have found a non-trivial divisor of n, and have successfully begun a factorization of n.

Well of course choosing integers b at random seems rather pointless, but perhaps we can nudge things in our favor. Choose an integer $a < n$ at random. If $1 < \gcd(a,n) < n$, we were very lucky. If not, then $\gcd(a,n) = 1$. Now if p is a prime with $p \mid n$, then since $\gcd(a,n) = 1$ we know $p \nmid n$ so $a^{p-1} \equiv 1 \pmod{p}$, that is $p \mid \gcd(a^{p-1}, n)$. Great, but we still don't know p. Yes, but if M is any integer so that $(p-1) \mid M$ we would also have $a^M = (a^{p-1})^{M/p-1} \equiv 1 \pmod{p}$, so that $p \mid \gcd(a^M - 1, n)$.

Now we get to the part where we hope there is a prime $p \mid n$ so that $p-1$ is the product of small primes to small powers. Choose a bound B and compute (a table of) prime powers less than or equal to B:

$$2^{e_2} \leq B, \qquad 3^{e_3} \leq B, \qquad 5^{e_5} \leq B, \qquad \cdots \qquad p_r^{e_r} \leq B.$$

For example with $B = 11$, we would have computed: $2^3, 3^2, 5^1, 7^1, 11^1$. Put $M = 2^{e_2}3^{e_3} \cdots p_r^{e_r}$, and compute $\gcd(a^M - 1, n)$. If $p-1$ is the product of small primes to small powers, then $p-1$ will divide such an M for B large enough, forcing $\gcd(a^M - 1, n)$ to be divisible by p. So the key to this method is to generate M and to compute $\gcd(a^M - 1, n)$. At first blush, this looks like it might be computationally intense, but it is not. We know that if $a \equiv b \pmod{n}$, then $\gcd(a,n) = \gcd(b,n)$, so we need not compute $a^M - 1$ exactly simply compute it modulo n, for which we already know of fast methods by expanding the exponent M in binary and then performing successive squarings.

The general algorithm: We are given an odd, composite number n, and an initial bound B.

(1) Find primes and prime powers $p_i^{a_i} \leq B$, $i = 1, 2, \ldots, r$. This can be done via known tables of primes, sieving, and other fast methods.
(2) Choose an integer $1 < a < n$ at random. If the $\gcd(a,n) > 1$, we have found a nontrivial factor. Otherwise, $\gcd(a,n) = 1$, and Euler's theorem applies.
(3) For $M = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$, compute $a^M - 1 \pmod{n}$.
(4) Test $d = \gcd(a^M - 1, n)$. If $1 < d < n$, we have succeeded, otherwise this iteration has failed.

After a couple of examples, we give options for how to proceed in the event of a failure.

Example 2.1. Let $n = 246082373$. Choose $a = 2$ and $B = 8$. Then $M = 2^3 \cdot 3 \cdot 5 \cdot 7$. We compute $\gcd(a^M - 1, n) = 1$ (using Euclid's algorithm). We have failed, so we increase B to 10. Now $M = 2^2 3^2 \cdot 5 \cdot 7 = 2520$. We find that $\gcd(a^M - 1, n) = 2521$ a prime! That is, $a^M = a^{p-1} \equiv 1 \pmod{p}$ (by Fermat) and so $p = 2521$ divides both n and $a^M - 1$. Using this divisor, we factor $n = 2521 \cdot 97613$, both of which are primes.

Example 2.2. $n = 2047 = 2^{11} - 1$. We choose $a = 2$ and $B = 10$, so $M = 2^2 3^2 \cdot 5 \cdot 7 = 2520$, and we find $\gcd(a^M - 1, n) = 1$, so we increase B to 11. Now $M = 2^2 3^2 \cdot 5 \cdot 7 \cdot 11$, and we find $\gcd(a^M - 1, n) = 2047 = n$, so increasing B can't help. We show this by observing that $\gcd(2^{11k} - 1, 2^{11} - 1) = \gcd(2^M - 1, n) = n$ for any k. Using the polynomial identity $x^k - 1 = (x - 1)(x^{k-1} + x^{k-2} + \cdots + x + 1)$, and substituting $x = 2^{11}$, we see $2^{11k} - 1 = (2^{11} - 1)(2^{11(k-1)} + \cdots + 1) \equiv 0 \pmod{2^{11} - 1}$, so our alternative is now to change the choice of a, to say $a = 3$. This type of failure occurs infrequently. Indeed [**1**], if n is divisible by at least two odd primes, the probability that a randomly chosen value for a will fail is less than one-half.

In summary, Pollard's $p - 1$ test works if there is a prime p dividing our composite number n so that $p - 1$ is the product of small primes to small powers, and depends on the fact that the group of units U_p has order $p - 1$. Actually, U_p is cyclic of order $p - 1$ (useful for El Gamal), though here we are simply using Lagrange's theorem (Fermat) to say $a^{|U_p|} = a^{p-1} = 1$ in U_p.

3. Elliptic Curve Factorization

The elliptic curve method of factorization (ECM) was developed by Henrik Lenstra, and is currently at the cutting edge in factorization techniques. In a real sense it is a natural and broad generalization of Pollard's $p - 1$ method, so we start the introduction with a comparison of the mechanics of the $p - 1$ and the ECM as algorithms to factor a composite integer n.

Pollard's $p - 1$	versus	Lenstra's ECM
U_p	$\longleftrightarrow$	$E(\mathbb{Z}_p)$
$(a \in U_p)$	$\longleftrightarrow$	$(P \in E(\mathbb{Z}_p))$
$a^M \equiv 1 \pmod{p}$ in U_p (if $p - 1 = \#U_p \mid M$)	$\longleftrightarrow$	$kP = \mathcal{O}$ in $E(\mathbb{Z}_p)$ (if $\#E(\mathbb{Z}_p) \mid k$)
Here M is the product of small primes to small powers	$\longleftrightarrow$	Here k is the product of small primes to small powers
We win if there is a prime $p \mid n$ with $\#U_p$ the product of small primes to small powers	$\longleftrightarrow$	We win if there is a prime $p \mid n$ with $\#E(\mathbb{Z}_p)$ the product of small primes to small powers

The big difference between these methods is that in the case of Pollard's $p - 1$, for each prime p, there is only one group associated to p, namely, U_p to exploit. In the case of the ECM, for each p, we have an enormous supply of elliptic curves having a broad spectrum of orders to utilize.

First, we should discuss how to compute kP, that is, the point P on the elliptic curve added to itself k times. Just as with exponentiation, this can be done effectively by using the binary expansion of k: $k = k_0 + k_1 2^1 + k_2 2^2 + \cdots + k_r 2^r$. We know how to add points on an elliptic curve; doubling a point $P \mapsto 2P$ is done via the tangent line, while if $P \neq Q$ the sum $P + Q$ is determined using the line through P and Q. So we precompute:

$$\begin{aligned} P_0 &= P \\ P_1 &= 2P_0 = 2P \\ P_2 &= 2P_1 = 2^2 P \\ P_3 &= 2P_2 = 2^3 P \\ &\vdots \\ P_r &= 2P_{r-1} = 2^r P \end{aligned}$$

Then $kP = k_0 P_0 + k_1 P_1 + \cdots + k_r P_r$. Note that the k_i are either 0 or 1, so we are just adding up the points P_i where k_i is nonzero.

To implement the procedure above, we review the group law for points on the curve $E_{a,b} : y^2 = x^3 + ax + b$.

Summary of Group Law: Let $\mathcal{O} = [0, 1, 0]$ and $P_i = [x_i, y_i, 1]$.

(1) $-\mathcal{O} = [0, -1, 0] = [0, 1, 0] = \mathcal{O}$.
(2) $-P_i = [x_i, -y_i, 1]$.
(3) $P_i + \mathcal{O} = \mathcal{O} + P_i = P_i$.
(4) $P_2 = -P_1$ iff $P_1 + P_2 = \mathcal{O}$,
(5) If $P_2 \neq -P_1$, then $P_1 + P_2 = P_3 = [x_3, y_3, 1]$, where

$$x_3 = m^2 - x_1 - x_2,\ -y_3 = m(x_3 - x_1) + y_1 \text{ and where } m = \begin{cases} \frac{y_2 - y_1}{x_2 - x_1} & \text{if } x_1 \neq x_2 \\ \frac{3x_1^2 + a}{2y_1} & \text{if } x_1 = x_2. \end{cases}$$

Let's take a second look at Example 4.1.

EXAMPLE 3.1. Consider the example [**6**] of the cubic curve $y^2 = x^3 + x + 1$ over the field $\mathbb{Z}_5$. The discriminant is $4a^3 + 27b^2 = 4 + 27 = 31 \not\equiv 0 \pmod 5$, so this is an elliptic curve over $\mathbb{Z}_5$. We check directly that

$$E(\mathbb{Z}_5) = \{\mathcal{O}, (0, \pm 1), (2, \pm 1), (3, \pm 1), (4, \pm 1)\},$$

and we have seen that $E(\mathbb{Z}_5) \cong \mathbb{Z}_9$ is a cyclic group of order 9, thus $9P = \mathcal{O}$ for any point P on the curve.

Consider the point $P = (0,1) = [0,1,1]$ on the curve. Write $k = 9 = 2^0 + 2^3$ in its binary expansion: $9 = (1001)_2$, so $9P = P + 8P$ and we compute $8P$ by successive doublings.

To double the point P we use the slope of the tangent line in our formulas above:

$$m = \frac{3x_2^2 + a}{2y_1} = \frac{3 \cdot 0 + 1}{2 \cdot 1} = \frac{1}{2} = 3 \text{ in } \mathbb{Z}_5.$$

Thus, $2P = [x_3, y_3, 1] = [4, 2, 1]$ (by our formulas). Similarly, we compute:

$$\begin{aligned} P &= [0,1,1], \\ 2P &= [4,2,1], \\ 4P &= [3,-1,1], \\ 8P &= [0,-1,1]. \end{aligned}$$

Of course knowing that $9P = P + 8P = \mathcal{O}$, we could have written down $8P$ without any computation since $P + Q = \mathcal{O}$ iff $Q = -P$, thus $8P = -P$ which we easily verify from the addition law for points on the elliptic curve.

A strict analogy of Pollard's $p-1$ would have us compute the gcd of k and n (the integer to be factored), but Lenstra's method is even more clever. One computes $kP = k_0P_0 + k_1P_1 + \cdots + k_rP_r$ as before with the binary expansion of k, but for each sum (or doubling for that matter), one must compute a slope, $m = (y_2 - y_1)/(x_2 - x_1)$ =r $m = (3x_1^2 + a)/2y_1$. Now in reality we are working with an elliptic "pseudocurve", $E_{a,b}(\mathbb{Z}_n) : y^2 = x^3 + ax + b$ in which we are treating this curve as an elliptic curve over $\mathbb{Z}_n$. When n is not a prime not all nonzero elements of $\mathbb{Z}_n$ have multiplicative inverses (indeed only those relatively prime to n, and so in computing the inverses associated to the slopes m, we can detect a factor of n by noting the failure of a denominator in m have an inverse modulo n, so at each multiplication we check $\gcd(x_2 - x_1, n)$ or $\gcd(2y_1, n)$. Producing a nontrivial gcd gives us a factor of n.

Lenstra's ECM algorithm. Given a composite integer n to factor, we

1. Check $\gcd(n, 6) = 1$ (2's and 3's make life with elliptic curves more difficult)
2. Check that n is not a perfect power, i.e., $n \neq m^k$ for some m and k. This is easy and quick to do. Just check (using real-valued functions) that none of $\sqrt{n}$, $\sqrt[3]{n}$, ..., $\sqrt[\ell]{n}$ are integers for $\ell = \lceil \ln n / \ln 2 \rceil$ (guarantees that $\sqrt[\ell]{n} < 2$).
3. Choose a bound B (say ~ 10000).
4. Choose a curve $E_{a,b}(\mathbb{Z}_n) : y^2 = x^3 + ax + b$ and a point $P = [x, y, 1]$ on the curve as follows:
   (a) Choose random integers $x, y, a \in [0, n-1]$.
   (b) Compute $b \equiv (y^2 - x^3 - ax) \pmod{n}$.

(c) Compute $d = \gcd(4a^3 + 27b^2, n)$. If $d = n$, start over choosing a new x, y, a. If $1 < d < n$, then d is a proper factor of n, and we have succeeded. Otherwise $d = 1$ which means we have an elliptic pseudocurve over $\mathbb{Z}_n$ (in particular an honest elliptic curve over $\mathbb{Z}_p$ for any prime $p \mid n$), and a point $P = [x, y, 1]$ on that curve.

(5) Compute highest prime powers less than or equal to the bound B:
$2^{a_2}, 3^{a_3}, \ldots, p_r^{a_r} \leq B$.

(6) Technically, we are hoping that if $k = 2^{a_2}3^{a_3} \cdots p_r^{a_r}$, that $kP = \mathcal{O}$, but we will actually compute kP in stages hoping for a failure anywhere along the way, for example:

$$\begin{gathered} P \mapsto 2P \mapsto 4P \mapsto \cdots \mapsto 2^{a_2}P \mapsto 3 \cdot 2^{a_2}P \mapsto 3^2 \cdot 2^{a_2}P \mapsto \cdots 3^{a_3} \cdot 2^{a_2}P \mapsto \cdots \\ \cdots \mapsto p_r \cdot (p_{r-1}^{a_{r-1}} \cdots 3^{a_3} \cdot 2^{a_2}P) \mapsto \cdots \mapsto p_r^{a_r} \cdot (p_{r-1}^{a_{r-1}} \cdots 3^{a_3} \cdot 2^{a_2}P) = kP. \end{gathered}$$

At each addition or doubling, we are looking to find a slope which cannot be computed by failure of the gcd of the denominator and n to be one. If the gcd is one, we continue the arithmetic; if the gcd is a proper divisor of n, we return the factor; if the gcd is n we can increase the bound B or try another curve.

It turns out that the computational complexity of the ECM is related to the size of the smallest prime factor which divides n, and not very much to n itself. This means ECM can be effective in finding a divisor of enormous composites (those with at least one not-so-large factor), but is worst at factoring RSA composites which are the product of two primes roughly the same size.

CHAPTER 8

Deeper Results and Concluding thoughts

Understanding the ECM of the last section was the primary goal of the course. All things beyond that were icing on the cake, though the first was highly desirable because it brings closure to an early and pervasive topic in the course (congruent numbers) and points the way for even further investigations.

1. The congruent number problem and Tunnell's solution

The main reference for this part is Koblitz's wonderfully written text [**4**].

Earlier in the course we defined congruent numbers, and established a relationship between Pythagorean triples and congruent numbers. Moreover, by parametrizing the rational points on the unit circle, we were able to list all Pythagorean triples, and if we let the listing of Pythagorean triples continue forever, eventually all congruent numbers would be listed. The problem is that even if one knew a number was a congruent number, there was no way of telling how long one would wait before it was listed corresponding to some Pythagorean triple.

All that changed in 1983 with Tunnell's elegant answer to the congruent number problem. What is most elegant about the answer is that it is just as easy to understand as the problem itself, that is it would have been a perfectly acceptable answer to the Greeks. The mathematics which underlies that answer, however, is quite deep, yet topics we shall at least broach.

Tunnell's answer is given by the following theorem (see [**4**]) which determines whether n is a congruent number by comparing the representation numbers of two ternary quadratic forms.

Theorem (Tunnell). *Let n be a squarefree positive integer.*

(1) Suppose that n is a congruent number.
If n is odd, then

$$\#\{(x,y,z)\in\mathbb{Z}^3 \mid n = 2x^2+y^2+32z^2\} = \frac{1}{2}\#\{(x,y,z)\in\mathbb{Z}^3 \mid n = 2x^2+y^2+8z^2\}.$$

If n is even, then

$$\#\{(x,y,z)\in\mathbb{Z}^3 \mid \frac{n}{2} = 2x^2+y^2+32z^2\} = \frac{1}{2}\#\{(x,y,z)\in\mathbb{Z}^3 \mid \frac{n}{2} = 2x^2+y^2+8z^2\}.$$

(2) Conversely, if the weak Birch-Swinnerton-Dyer conjecture is true for elliptic curves of the form $E_n : y^2 = x^3 - n^2x$, then these equalities of cardinalities implies that n is a congruent number.

For example on page 5 of [**4**], the author lists an example of Zagier which shows that 157 is a congruent number. For example the hypotenuse is a rational number which (in reduced form) has a denominator with 45 digits. Clearly, this would not be so easy to find. On the other hand, we consider Tunnell's theorem.

The number 157 is odd and squarefree, so we need only verify the cardinalities

$$\#\{(x,y,z) \in \mathbb{Z}^3 \mid 157 = 2x^2 + y^2 + 32z^2\} = \frac{1}{2}\#\{(x,y,z) \in \mathbb{Z}^3 \mid 157 = 2x^2 + y^2 + 8z^2\}.$$

We claim that the cardinality of both sets is zero, that is there are no solutions. Note that if $157 = 2x^2 + y^2 + 8z^2$ has no solutions, than neither can $157 = 2x^2 + y^2 + 32z^2$ since $32z^2 = 8(2z)^2$. Consider the equality $2x^2 + y^2 + 8z^2 = 157$ as a congruence modulo 8: we obtain $2x^2 + y^2 \equiv 5 \pmod 8$ which implies that y is odd, hence $y^2 \equiv 1 \pmod 8$. It follows that $2x^2 \equiv 0, 2 \pmod 8$, so that $2x^2 + y^2 + 8z^2 \equiv 1, 3 \pmod 8$, so there can be no solutions. But both sets have the same (zero) cardinality, so by the theorem, 157 is a congruent number.

Oh yes, what about this Birch-Swinnerton-Dyer conjecture, and where did elliptic curves come in? That will take us a bit longer to explain.

We follow Koblitz [**4**] here. We begin with a series of propositions to connect congruent numbers to points on elliptic curves.

Proposition. *Let $n \geq 1$ be a squarefree integer. Let $X, Y, Z \in \mathbb{Q}$ with $X < Y < Z$. There is a one-to-one correspondence between right triangles with sides X, Y and hypotenuse Z having area n, and rational numbers x so that $x, x+n, x-n$ are all squares in $\mathbb{Q}$. The correspondence is given by:*

$$X, Y, Z \mapsto x = (Z/2)^2$$
$$x \mapsto X = \sqrt{x+n} - \sqrt{x-n}, \quad Y = \sqrt{x+n} + \sqrt{x-n}, \quad Z = 2\sqrt{x}$$

Proof. Let's first see that the numbers do what is claimed. Given X, Y, Z with $X^2 + Y^2 = Z^2$ and $XY/2 = n$ we see that

$$\begin{aligned} X^2 + Y^2 \pm 4\frac{1}{2}XY &= Z^2 \pm 4n, \text{ or} \\ (X \pm Y)^2 &= Z^2 \pm 4n, \text{ or} \\ \left(\frac{X \pm Y}{2}\right)^2 &= \left(\frac{Z}{2}\right)^2 \pm n. \end{aligned}$$

So $x = (Z/2)^2$ is obviously a square, and hence so is $x \pm n = (Z/2)^2 \pm n = ((X \pm Y)/2)^2$.

Conversely, given $x \in \mathbb{Q}$ with $x, x \pm n$ all squares, we put $X = \sqrt{x+n} - \sqrt{x-n}$, $Y = \sqrt{x+n} + \sqrt{x-n}$, and $Z = 2\sqrt{x}$, all of which are now rational numbers by the assumption. We see that

$$\frac{1}{2}XY = \frac{1}{2}(x+n-(x-n) = n, \text{ and}$$
$$X^2 + Y^2 = 2(x+n+x-n) = 4x = Z^2.$$

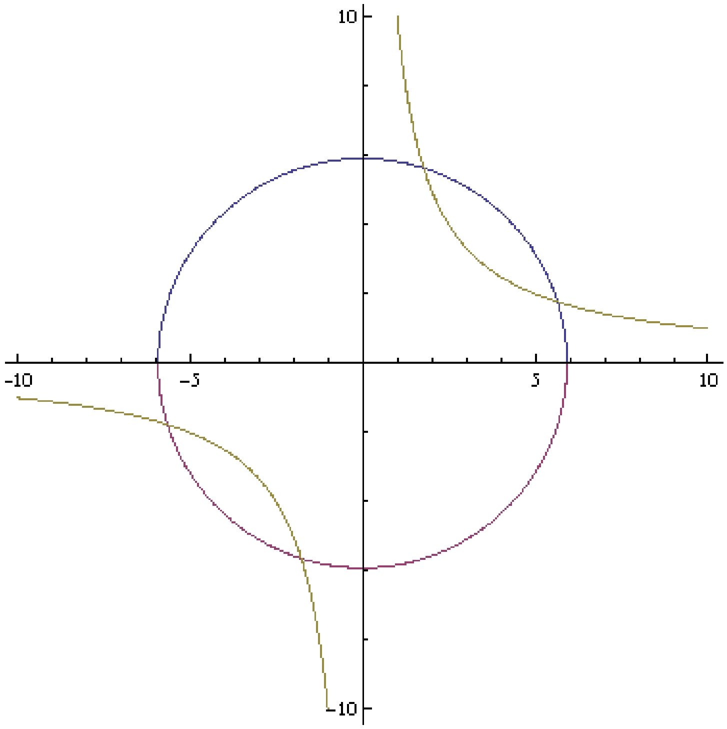
So X, Y, Z are the sides of a rational right triangle with area n. It is trivial to check that $X < Y < Z$.

Now we see that there is a one-to-one correspondence. Let $x \longleftrightarrow X, Y, Z$ and $x' \longleftrightarrow X', Y', Z'$

Suppose that $(X, Y, Z) = (X', Y', Z')$. The fact that $Z = 2\sqrt{x} = Z' = 2\sqrt{x'}$ implies $x = x'$ (since both are positive). Conversely, suppose that $x = x'$. Then $Z = 2\sqrt{x} = Z' = 2\sqrt{x'}$ and hence

$$Z'^2 = X'^2 + Y'^2 = X^2 + Y^2 = Z^2, \text{ and } \frac{1}{2}XY = n = \frac{1}{2}X'Y'.$$

Geometrically, we are looking at the intersection of the circle $X^2 + Y^2 = Z^2$ (with $Z = Z'$ fixed), and the hyperbola $\frac{1}{2}XY = n$. The typical situation is pictured below:



So there are only four possible points (X, Y) which work, and the constraints $0 < X < Y$ mean there is only one point, and the proof is complete. □

We saw above that $\left(\frac{X\pm Y}{2}\right)^2 = \left(\frac{Z}{2}\right)^2 \pm n$. Multiplying the two expressions together yields $\left(\frac{X^2-Y^2}{4}\right)^2 = \left(\frac{Z}{2}\right)^4 - n^2$ which says the curve $u^4 - n^2 = v^2$ has a rational solution $u = (X^2 - Y^2)/4$, $v = (Z/2)$. Multiplying the equation if u, v by u^2 yields $u^6 - n^2u^2 = (uv)^2$, so putting

$$x = (Z/2)^2 = u^2, \quad y = (uv) = (X^2 - Y^2)Z/8,$$

we have a rational point on the elliptic curve $y^2 = x^3 - n^2x$. Conversely, we have

Proposition. *Let (x, y) be a rational point on the elliptic curve $y^2 = x^3 - n^2x$ Suppose that x is a square and has even denominator. Then putting*

$$X = \sqrt{x+n} - \sqrt{x-n}, \quad Y = \sqrt{x+n} + \sqrt{x-n}, \quad Z = 2\sqrt{x}$$

produces a rational right triangle with area n.

Denote by E_n the elliptic curve $y^2 = x^3 - n^2x$. Critical to moving forward is Mordell's important theorem describing the structure of the group of rational points on E_n, denoted $E_n(\mathbb{Q})$.

Theorem (Mordell). $E_n(\mathbb{Q}) \cong \mathbb{Z}^r \oplus E_n(\mathbb{Q})_{tor}$.

Here $E_n(\mathbb{Q})_{tor}$ is the *torsion* subgroup, that is the point in $E_n(\mathbb{Q})$ having finite order. The integer $r \geq 0$ is called the rank of the elliptic curve and $r > 0$ iff there are infinitely many rational points on E_n. More precisely, the theorem says there are r points $P_1, \ldots, P_r \in E_n(\mathbb{Q})$ so that for any point $P \in E_n(\mathbb{Q})$, P can be written uniquely as $P = k_1P_1 + k_2P_2 + \cdots + k_rP_r + Q$ for (unique) integers $k_1, \ldots, k_r$ and a unique torsion point Q.

It is clear from the graph of the curve that the points $[-n, 0, 1], [0, 0, 1], [n, 0, 1]$ all have order 2. This means that these three points (together with the identity $\mathcal{O}$) are all elements of $E_n(\mathbb{Q})_{tor}$. On the other hand we have the theorem:

Theorem. *$\#E_n(\mathbb{Q})_{tor} = 4$ for all squarefree positive n.*

Knowing the fundamental theorem of finite abelian groups, we then deduce that

Corollary. $E_n(\mathbb{Q})_{tor} = \{\mathcal{O}, [-n, 0, 1], [0, 0, 1], [n, 0, 1]\} \cong \mathbb{Z}_2 \times \mathbb{Z}_2$.

Finally, we come to the theorem which ties these ideas together:

Theorem. *A positive, squarefree integer n is a congruent number if and only if $E_n(\mathbb{Q})$ has positive rank, which is to say if and only if it has infinitely many rational points.*

Proof. We sketch the proof. If n is a congruent number, then we have seen there exists a point $(x, y) \in E_n(\mathbb{Q})$ with x a positive square. By inspection, such a point is not in the torsion subgroup, so is a point of infinite order. Conversely, if P is a point of infinite order in $E_n(\mathbb{Q})$, then using our doubling formula, we easily check that $2P$ has x-coordinate a square with even denominator which by previous work shows n is a congruent number. □

This ends the so-called easy part of Tunnell's proof and occupies only the first chapter of Koblitz's book [**4**]. We go a bit further to describe the Birch-Swinnerton-Dyer conjecture, but we do so more to advertise the role and independent interest of complex analysis than to just define the analytic objects involved with the B-S-D conjecture.

2. A digression on functions of a complex variable

At first blush, one might presume there to be little difference in studying differentiable functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $f : \mathbb{C} \to \mathbb{C}$, but there is, and the differences are dramatic. We point out three important distinctions.

First, if a functions $f : \mathbb{C} \to \mathbb{C}$ has a continuous first derivative, it is infinitely differentiable and indeed can be expressed as a power series. The same is certainly not true even for functions $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = x^{5/3}$ shows. Indeed this remarkable property is related to another which says that if one knows the values of the differentiable function $f : \mathbb{C} to \mathbb{C}$ on the boundary of a nice region, then the values of f on the interior of the region are determined. Geometrically, this implies a certain rigidity. We can't see the graph of a function $f : \mathbb{C} \to \mathbb{C}$ (it lives in $\mathbb{C}^2 \cong \mathbb{R}^4$), but if we could and if the same were true of functions $f : \mathbb{R}^2 \to \mathbb{R}$ (whose graph is a surface), we would see the following. Imagine a function defined on the closed unit disk and which is zero on the boundary. Surely you could draw lots of surfaces like that. If that were an analytic function, there would only be one graph since the values on the boundary determine the values inside. Strange? Yes

Second, it is often the case the we have two definitions of a function (or more precisely) two functions whose definitions agree on a nice set. Then they agree everywhere they are both defined. First we give an example of how this does not happen for differentiable real-valued functions. Consider two functions $f, g : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} (x-1)^4 & \text{if } x \geq 1 \\ 0 & \text{otherwise} \end{cases} \qquad g(x) = \begin{cases} (x+1)^4 & \text{if } x \leq -1 \\ 0 & \text{otherwise.} \end{cases}$$

These are both smooth functions, defined on all of $\mathbb{R}$a whose values agree for all $x \in [-1, 1]$, but which are clearly not the same wherever both are defined.

This can't happen in the complex case, and this turns out to be very handy. Consider the Riemann zeta function

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

probably one of the most famous functions in all of mathematics. You studied this object in calculus when you worked with infinite series. You studied so-called p-series which have the form $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^p}$, and showed that these series converge when $p > 1$. Recall that $p = 1$

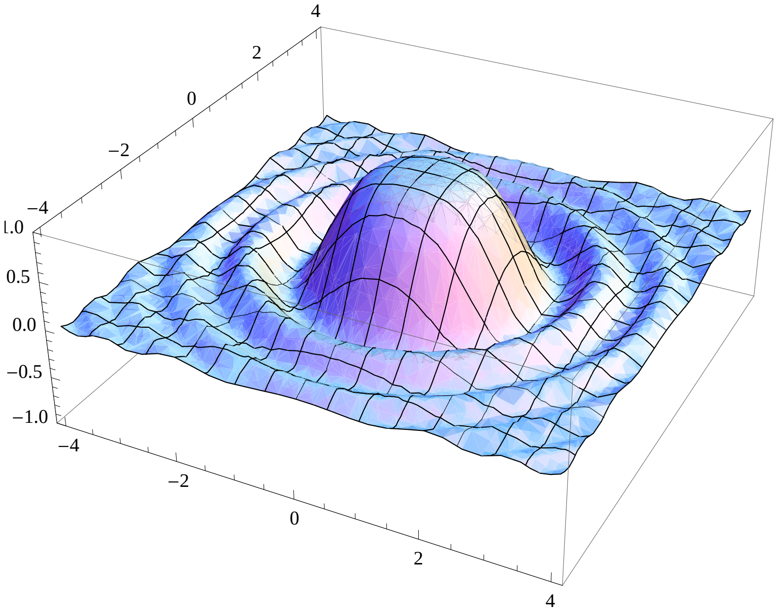
corresponds to the harmonic series which diverges, and for $p < 1$ the series diverges by comparison.

What the calculus result really says is that $\zeta(s)$ is a function whose domain in $\mathbb{R}$ is $(1, \infty)$. It doesn't take much more effort to look at series in $\mathbb{C}$ and we find that the actual domain of $\zeta(s)$ is the right half-plane $\Re(s) > 1$. Now here comes the rub. There is a famous conjecture (better than Fermat) which says that (except for some trivial cases) the function $\zeta(s)$ is zero only when $\Re(s) = 1/2$. This is the famous Riemann hypothesis. There is just one problem. The function isn't even defined where the hypothesis is telling us to look. That's where the identity theorem comes in. Suppose with a bit more math, you could define an analytic function $Z(s)$ whose domain was all of $\mathbb{C}$ except for the point $s = 1$, and for which $Z(x) = \zeta(x)$ for all real $x > 1$. Then $Z(s) = \zeta(s)$ for all complex points $\Re(s) > 1$ and $Z(s)$ defines what is called an analytic continuation of the zeta function $\zeta(s)$. The new function $Z(s)$ is the one to which the Riemann hypothesis refers.

The third distinction between real analytic and complex analytic functions is Louiville's theorem, which says that any function $f : \mathbb{C} \to \mathbb{C}$ which is analytic in all of $\mathbb{C}$ (called an entire function) and bounded must be a constant. This certainly doesn't happen for real-valued functions, e.g., $\sin x$, or even more complicated ones, such as $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = \begin{cases} \frac{\sin(x^2+y^2)}{x^2+y^2} & (x, y) \neq (0.0) \\ 1 & (x, y) = (0, 0). \end{cases}$$

The function is infinitely differentiable and bounded between -1 and 1, but obviously not constant.



3. Return to B-S-D

We start with the elliptic curve $E_n : y^2 = x^3 - n^2x$. Notice that while we have been interested in the set of rational points, $E_n(\mathbb{Q})$, it also makes sense to think about $E_n(\mathbb{Z}_p)$ at least for primes $p \nmid n$. Since over the finite field, the number of points is finite, we can count them and record the information as follows: For $p \nmid 2n$, let $E_n(\mathbb{Z}_p) = p + 1 - a_p$, the shape here influenced by Hasse's theorem. At any rate one forms the Hasse-Weil L-function:

$$L(E_n, s) = \prod_{p \nmid 2n} \left(\frac{1}{1 - a_p p^{-s} + p^{1-2s}} \right).$$

As mysterious as this looks, it is just a complex valued function, very much like the Riemann zeta function, which is defined on the half-plane $\Re(s) > 3/2$. As with the Riemann zeta function, $L(E_n, s)$ has an analytic continuation to the whole complex plane, and in particular is defined at $s = 1$. Since the function is analytic at $s = 1$ it makes sense to talk about its order of vanishing. For example, for real-valued functions, $f(x) = x^2$ is nonzero at $x = 1$ so has zero order of vanishing. $f(x) = (x-1)^r(x^2+3)$ has kth order vanishing. So the definition is write $L(E_n, s) = (s-1)^k g(s)$ where $g(1) \neq 0$. Then k is the order of vanishing.

One version of the B-S-D conjecture is that if $L(E_n, s) = (s-1)^k g(s)$ and $E_n(\mathbb{Q}) = \mathbb{Z}^r \oplus E_n(\mathbb{Q})_{tor}$, then $r = k$, that is the rank of the elliptic curve is the order of vanishing of its L-function.

Ok, we bring this all the way back to the congruent number problem. We knew that n was a congruent number depended upon $E_n(\mathbb{Q})$ having infinitely many rational points. This is the same as saying the rank r is positive, which is simply to say $L(E_n, 1) = 0$. Phew!

Actually half of this connection is known. The Coates-Wiles theorem says that if $r \geq 1$, then $L(E_n, 1) = 0$. The converse is (one version of the) B-S-D conjecture.

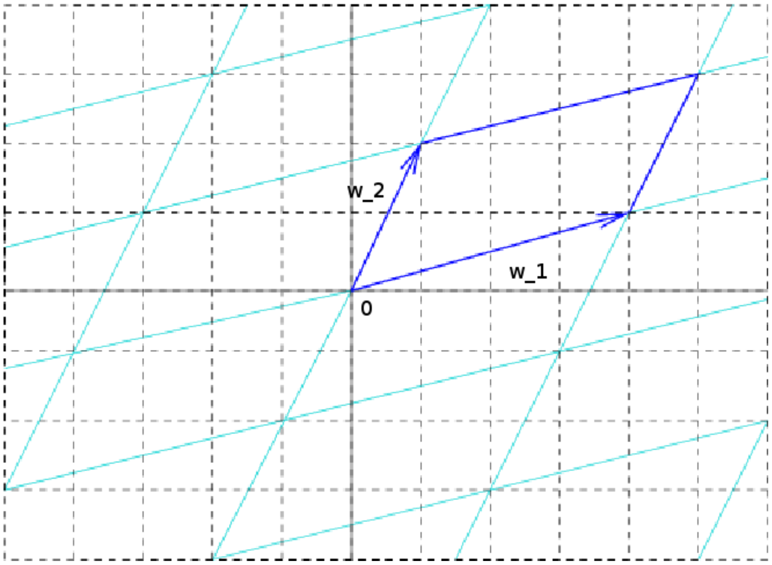
Finally connecting the vanishing of $L(E_n, s)$ at $s = 1$ to the formulas in Tunnell's theorem is where things *really* get exciting, but to talk about that we need modular forms and the Shimura lift.

CHAPTER 9

A closing topic

Here we give a sketch that an elliptic curve over $\mathbb{C}$ is a torus. In a sense this topic also will revisit many of the ideas we have developed in the course: equivalence relations, groups, projective curves, complex variables, and more. It is just for edification and amusement, so relax; it's not on the final.

Consider a piece of the complex plane shown below (modified slightly from `http://commons.wikimedia.org/wiki/File:Fundamental_parallelogram.png`) with the rectangular grid lines at the points $a + bi$ with $a, b \in \mathbb{Z}$.



Consider the vectors $\omega_1 = 4 + i$ and $\omega_2 = 1 + 2i$. These form two sides of what we will call the fundamental parallelogram Π. The vertex diagonally opposite the origin is $\omega_1 + \omega_2$. There are several things to observe about this image. First is that Π tiles the plane by translation; you can see the translated parallelograms in light blue. Second the vertices of all the blue parallelograms are precisely the set of points $\Lambda = \{a\omega_1 + b\omega_2 \mid a, b \in \mathbb{Z}\}$. More importantly, Λ is an abelian group under addition. What does that mean?

The set of points in the fundamental parallelogram $\Pi = \{a\omega_1 + b\omega_2 \mid 0 \leq a, b \leq 1\}$, so we have just blurred the distinction between the parallelogram and the region it bounds, and that's just fine.

Let's define a relation on the points in $\mathbb{C}$. We shall say $z_1 \sim z_2$ if and only if $z_2 - z_1 \in \Lambda$. It is trivial to check that this is reflexive, symmetric and transitive, hence an equivalence relation. One checks that every point in $\mathbb{C}$ is equivalent (can be translated by integer multiples of ω_1 and ω_2) to a point in Π. For example, $-4.5 + 26.3i \sim .5 + .3i \in \Pi$. Moreover, two points in Π are equivalent if and only if they are on the boundary: related by $z \sim z + \omega_i$. This has an important geometric (or really topological) interpretation. What we have said is that opposite sides of the parallelogram bounding Π should be identified. If you think of the parallelogram as a sheet of paper, when we identify two opposite edges, we roll the paper up into a cylinder gluing the two edges together. Now imagine the cylinder long and flexible. We could then fold the two ends up and glue them together forming a donut, mathematically a torus. It turns out that the torus is a group and as a group is isomorphic to the group of points on a complex elliptic curve.

First, let's identify the group. We shall define $\mathbb{C}/\Lambda$, read "$\mathbb{C}$ mod Λ", to be the set of equivalence classes under our equivalence relation defined above:

$$\mathbb{C}/\Lambda = \{[z] \mid z \in \mathbb{C}\}.$$

This is just like defining $\mathbb{Z}_n$ from the equivalence relation on $\mathbb{Z}$ with $a \sim b$ iff $a \equiv b \pmod{n}$. In particular, we can define a group law on $\mathbb{C}/\Lambda$ by defining $[z]+[w] = [z+w]$. The identity is $[0]$ and the inverse of $[z]$ is $[-z]$. Moreover, there is a one-to-one correspondence between the elements of this group and the points on the torus. Most people identify the two.

Now we need an elliptic curve ($y^2 =$cubic) and a map from the torus to the elliptic curve. We start in what appears a roundabout manner.

We all know that periodic functions on the real line have interesting properties, starting with sine and cosine and progressing to the theory of Fourier series which is a good deal more important than the Taylor series you studied in calculus. But for now, we shall settle for the statement that periodic functions are important.

What about in the complex plane. What kinds of functions exist which satisfy $f(z+\omega_1) = f(z)$ and $f(z + \omega_2) = f(z)$? These are called doubly-periodic functions. Naively, we might look for analytic. doubly-periodic functions, but this turns out to be boring for the following reason. The first important observation is that if f is doubly-periodic, then all the values $f(z)$ are determined by $z \in Pi$. Now even a continuous function on a closed bounded region like Pi (the fancy word in compact) achieves an absolute maximum and a minimum, so on all of $\mathbb{C}$ is bounded. So if in addition f is analytic, then it is a bounded, entire function, which by Louiville, must be constant. This is what we meant by boring.

If f were analytic, it would have a power series $\sum_{n=0}^{\infty} a_n z^n$, but if it is not, it can still have a series expansion. It's just that there may be some negative exponents: $\sum_{n=\mu}^{\infty} a_n z^n$ for $\mu < 0$. Functions like this are called meromorphic.

Now we define yet another object without foreshadowing, but not unexpectedly.

Let $\mathcal{E}_\Lambda = \{f : \mathbb{C} \to \mathbb{C} \mid f \text{ is meromorphic and } f(z+\lambda) = f(z) \text{ for all } \lambda \in \Lambda\}$. It is clear that constant functions are in $\mathcal{E}_\Lambda$, so the set is non-empty. If $f, g \in \mathcal{E}_\Lambda$, then so is $f \pm g$, $f * g$ and f/g as long as $g \neq 0$. But this makes $\mathcal{E}_\Lambda$ a field, called the field of elliptic functions with period lattice Λ. Something else is true:

PROPOSITION. *If $f \in \mathcal{E}_\Lambda$, then so is its derivative, f'.*

PROOF. We really need only show that f' is also periodic. From calculus, we observe:

$$f'(z+\lambda) = \lim_{h\to 0} \frac{f(z+\lambda+h) - f(z+\lambda)}{h} = \lim_{h\to 0} \frac{f(z+h) - f(z)}{h} = f'(z),$$

where at the second equality we made use that f was periodic (i.e., in $\mathcal{E}_\Lambda$). □

There is a great deal more one could say, but we're almost at the end, so we'll push on. An extremely important example of an elliptic function is the Weierstrass function $\wp(z)$. We skip its formal definition and simply say that is has a series expansion of the form:

$$\wp(z) = \frac{1}{z^2} + a_2 z^2 + a_4 z^4 + a_6 z^6 + \cdots .$$

Equally important is its derivative:

$$\wp'(z) = \frac{-2}{z^3} + 2a_2 z + 4a_4 z^3 + 6a_6 z^5 + \cdots .$$

We consider $\wp'(z)^2$ and $\wp(z)^3$ and compare:

$$\begin{aligned} \wp'(z)^2 &= \frac{4}{z^6} - \frac{8a_2}{z^2} - 16a_4 + z^2(\cdots) \in \mathcal{E}_\Lambda \\ \wp(z)^3 &= \frac{1}{z^6} + \frac{3a_2}{z^2} + 3a_4 + z^2(\cdots) \in \mathcal{E}_\Lambda. \end{aligned}$$

We compute:

$$\begin{aligned} \wp'(z)^2 - 4\wp(z)^3 &= \frac{-20a_2}{z^2} - 28a_4 + z^2(\cdots), \text{ so} \\ \wp'(z)^2 - 4\wp(z)^3 + 20a_2\wp(z) &= -28a_4 + z^2(\cdots) \in \mathcal{E}_\Lambda. \end{aligned}$$

So what is the point of this? Well both sides of the last equation are elliptic functions in $\mathcal{E}_\Lambda$, but from the right hand side we see that the function is actually analytic, and by Louiville, analytic elliptic functions are constant, so the right hand side is just $-28a_4$. Putting this all together, we see that

$$\wp'(z)^2 = 4\wp(z)^3 - 20a_2\wp(z) - 28a_4.$$

Still don't see it coming? Too shell-shocked? Ok, let $x = \wp(z)$ and $y = \wp'(z)$. Then $y^2 = 4x^3 - 20a_2x - 28a_4$. The point $9x, y)$ is a point on an elliptic curve!

We define a function $\Phi : \mathbb{C}/\Lambda \to E : y^2 = 4x^3 - 20a_2x - 28a_4$ by

$$\Phi(z) = \begin{cases} [\wp(z), \wp'[z], 1] & z \neq 0, \\ [0, 1, 0] & z = 0. \end{cases}$$

ϕ is an isomorphism of groups. If $z_1 \mapsto P_1 = [\wp(z_1), \wp'(z_1), 1]$ and $z_2 \mapsto P_2 = [\wp(z_2), \wp'(z_2), 1]$, then $z_1 + z_2 \mapsto P_1 + P_2$ (as we would define the sum of points on the elliptic curve), which equals $[\wp(z_1 + z_2), \wp'(z_1 + z_2), 1]$.

Bibliography

1. Richard Crandall and Carl Pomerance, *Prime numbers*, second ed., Springer, New York, 2005, A computational perspective. MR MR2156291 (2006a:11005)
2. Gareth A. Jones and J. Mary Jones, *Elementary number theory*, Springer Undergraduate Mathematics Series, Springer-Verlag London Ltd., London, 1998. MR MR1610533 (2000b:11002)
3. Ann Hibner Koblitz, Neal Koblitz, and Alfred Menezes, *Elliptic curve cryptography: the serpentine course of a paradigm shift*, J. Number Theory **131** (2011), no. 5, 781–814. MR 2772472 (2012b:14052)
4. Neal Koblitz, *Introduction to elliptic curves and modular forms*, Graduate Texts in Mathematics, vol. 97, Springer-Verlag, New York, 1984. MR MR766911 (86c:11040)
5. Kenneth H. Rosen, *Elementary number theory and its applications*, fifth ed., Addison-Wesley, Reading, MA, 2005.
6. Joseph H. Silverman and John Tate, *Rational points on elliptic curves*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1992. MR MR1171452 (93g:11003)