Dwight School
New York, New York

Group Theory and Modern Cryptography

To what extent does Group Theory provide a mathematical backdrop for modern public-key cryptography? 340923 words

Mathematics

Author : Jacob Bruner
Advisor: Daniel Bjelis

Contents

Foreword

This paper is for research purposes only. Although the information presented is as close to accurate as possible, one should never implement a cryptographic system themselves unless they know exactly what they're doing. **I am not liable for any damages caused in testing/creating/implementing any cryptographic protocol referenced in this paper.** Cryptography's greatest weakness is human-error, and virtually *all* breaks in cryptography result from poor implementation.[1]

1 Introduction and Aim

1.1 Topic and Research Question

1.2 Motivating Problem

It's undeniable that our modern-day world is reliant on cryptography. Every time a phone sends a text, a browser connects to a server, an email gets sent off, a monetary transaction is made, and much much more, our devices are, unbeknownst to us, performing many hundreds of math operations to ensure our data are 'encrypted.' But what does 'encryption' mean? Let's introduce some definitions. 'Encryption' is the process of disguising a message to be, loosely speaking, hidden to all *except* the intended recipient. This is the process of converting a 'plaintext' message into a jumbled 'ciphertext', which can be readily shared without risk of the sensitive message leaking. Converting a plaintext message (typically a string/list of characters) into a ciphertext is known as an 'enciphering' or 'encrypting' transformation. Likewise the reverse operation of recovering the plaintext message from a ciphertext is known as the *deciphering transformation.*[2] If we denote the plain and cyphertext $\mathcal{P}$ and $\mathcal{C}$ respectively and the enciphering map f and its inverse f^{-1} we obtain the following diagram:

$$\mathcal{P} \xrightarrow{f} \mathcal{C} \xrightarrow{f^{-1}} \mathcal{P}$$

1.3 The Fundamentals of Symmetric Cryptography

The intuitive way to implement this cryptosystem has the two transacting parties agree upon the nature of the map f in secret, beforehand. This might mean meeting up with a friend in-person to establish the common secret, f and f^{-1}, so that you could encrypt and decrypt each other's emails—fending off any prying eyes accessing their emails. This type of system has a special name, *"Symmetric Key Cryptography"*, reflecting the fact that both parties

[1] **thimbleby**.
[2] **koblitz**.

have the same shared secret (foreshadowing). Important historical examples include the 'Caeser Cipher' (supposedly invented by *the* Julius Caeser), where f is a shift operation that maps each letter to a new one a number of places ahead or behind. For instance, the two parties might decide beforehand the map f: *shift each character forward in the alphabet 3 letters*, implying the inverse map f^{-1}: *shift each character back 3 letters*. This might look like so: [0]

$$\mathcal{P} \sim A, B, C, D, E, F, ...Y, Z$$
$$\mathcal{C} \sim D, E, F, G, H, I, ...B, C$$

And, for example, if you wanted to encrypt the message $\mathcal{P}$ = "HELLO", you would obtain the ciphertext $\mathcal{C}$ = "KHOOR" which you would promptly send off for your friend to decode with the inverse, subtract-three-letters map. Note that, even in this simple example, repeated letters, word length, and other syntactic information provide a lot of information about the nature of the plaintext. Although this is a trivial example, schemes which, loosely speaking, encode any information about the syntax (or related notions) of the plaintext are usually highly vulnerable to a technique called *differential cryptanalysis*,which, at a high level, measures the change in cyphertext given a change in input, primarily targeting published symmetric-key protocols.[3]

To see what a caeser cipher looks like mathematically, we start encoding each letter as a number from 0, which is A, to 25, which is Z. Of course this depends on what alphabet one uses, if one chooses to include spaces, numbers, punctuation etc. We can represent the operation that takes a letter and maps it n places ahead with addition. Importantly, this operation must 'wrap around' back to zero if you try to exceed 'z' in the alphabet. This process, known as *'modular arithmetic'*, is like circling a clock, where after reaching twelve, the hour hand wrap back around to 1, but we start from 0 instead of 1. So, in our case, shifting 'z' by 3 letters looks like so: $25 + 1 \bmod 26 \equiv 0$, which reads "25 plus 1 *is congruent to* 0 *modulo* (or *mod*) 26." With this in mind, we obtain for each letter $p \in \mathcal{P}$:

$$f(p) = p + n \bmod 26$$

Representing a shifting of each letter in the plaintext by n places in the alphabet. Now clearly this isn't a very sophisticated cryptographic scheme...

[0] The informed reader will notice that Ceaser Ciphers are really just linear transformations in disguise, provoking thought into whether more advanced techniques could be used. For instance, we could consider groups of two letters, called *digraphs*, resulting in f being a 2x2 invertible matrix, encoding the shift in a two-by-two matrix. Similarly, we could consider invertible *affine* transformations of the form $f : \vec{x} \to A\vec{x} + \vec{b}$, where A encodes a certain scaling factor.

[3] **koblitz**.

For instance, performing a frequency analysis and comparing the most commonly occurring letters to those of the English alphabet, easily breaking these types of cyphers, broadly referred to as 'substitution ciphers.'[4] Modern schemes typically employ more resistant techniques, namely where changing just one letter of plaintext often yields a completely different ciphertext, making modern techniques all but impervious to frequency analysis. (Unless someone *really* screwed up an implementation)

For instance, AES encryption, part of the modern web standard, is an example of a '*block substitution cypher*,' which are beyond the scope of this paper. At a high level, it combines techniques similar to our Caesar Cipher with certain affine (think scaling and transforming) maps performed on blocks of plaintext. It's certainly more complicated than that, but essentially boils down to our system plus some advanced techniques making it resistant to cryptanalysis. (Like permutations, combinations, text look-up-tables and more.) In general, Symmetric-key cryptography (*viz.* predetermined, shared secret) is well understood. For instance, Claude Shannon (*the* definitive father of information theory) proved mathematically that the so-called *'one-time-pad'* encryption technique was completely and utterly unbreakable. In the general case, he showed that, if a random-generated key is at least as long as the plaintext (at least specifying a unique, random[0] character for each plaintext character and in the same alphabet), then performing a caeser cypher shift on each *individual* plaintext character by the value specified by the random character yields a cryptosystem that is mathematically impenetrable. Or equivalantly, performing a random modular addition on each character of the plaintext, with the shared secret being the sequence of random shifts.[5] Although modern systems seek smaller key sizes for performance reasons, this worst-case scenario should demonstrate the strength of symmetric-key algorithms in general. The 'one-time-pad' gets its name from its use in WWII when the KGB would distribute palm-sized pads with these one-time-keys and a table to ease in conversion. Such pads were often made of flammable materials to be burned with no trace.[6]

1.4 The 'Public-Key' Paradox

In the modern world, it's impractical to require that every shared secret be determined ahead of time. If a user wants to connect to a twitter server

[4] **koblitz**.

[0] There is a paramount distinction between *random* and *psuedorandom*, the latter of which is vastly easier to implement on a computer. True randomness has to be derived from a non-computer source (for the most part). Commonly implemented approximates include taking the least-significant-digit of a mouse position, or the frequency of keypresses, etc. A one-time-pad–like scheme generated from a psuedorandom source *is* breakable, although usually not without some effort.

[5] **claude**.

[6] **lewand**.

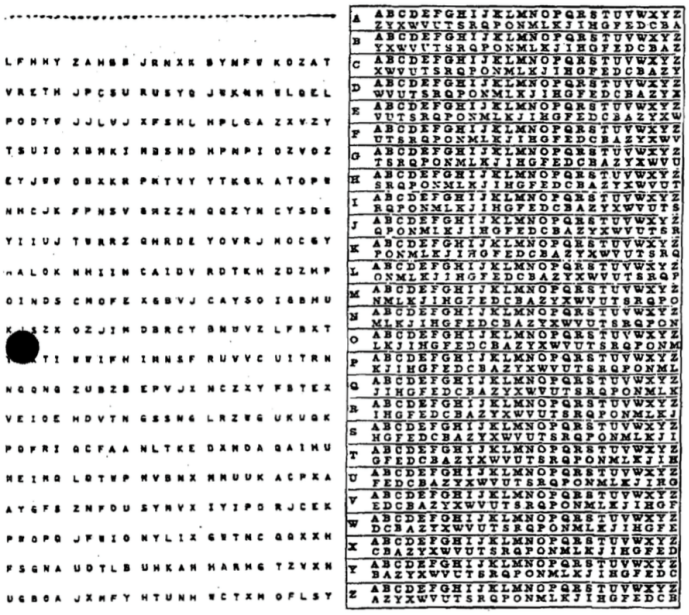


Figure 1: Format of a one-time-pad used by the NSA[7]

over a secure connection, how could symmetric-key encryption be employed? More generally, if two computers want to establish a connection for the first time, is there any way they could do so in an encrypted matter? The intuitive answer might be no, since how could you pass (or otherwise determine) a shared secret without any man-in-the-middle being able to obtain that same key. But this defys the ubiquity of encryption on the internet—every time I connect to a unfamiliar website, I still see a padlock on my browser. How can this be?

Consider the intuitive fact that some operations are more difficult to do in reverse than fowards—certain 'one-way functions.' For instance, if I mixed two different-colored paints together and asked whether you, *a priori*, could deduce the two initial colors given the end result, would you be able to? Although its quite easy to check whether any two colors combine to match the end result, there isn't an easy operation that takes the end result and returns the initial colors. This happens to be true for a number of operations.

[I have yet to finish this section.]

2 What is 'Public-Key' Cryptography

"In applied contexts, the terms "easy" and "hard" are usually interpreted relative to some specific computing entity; typically "cheap enough for the

legitimate users" and "prohibitively expensive for any malicious agents"."

The first protocol developed to address this motivating problem was RSA encryption. Leveraging intuitive properties of numbers, RSA establishes our idea of 'one-way operations' using simple multiplication of large, highly prime (minimal divisors), numbers.

0

[I have yet to finish this chapter. This will detail Diffie-Hellman key exchange over finite fields, discussing eucliean algorithm/bezouts id to compute inverses. I will then hint towards group theory before transitioning to the introduction.]

3 How can we formalize this: An Introduction to Groups

[a word to whoever is reading: I wrote this section before the preceeding sections, so it lacks many critical references to what I have been discussing throughout the paper. In the case of linear or affine cryptosystems, they correspond highly to additive groups $\mathbb{Z}/n\mathbb{Z}$ (n not necessarily prime), and, I believe, direct sums of such (for digraphs, trigraphs, etc). Similarly, Diffie Hellman is secretly group theory in disguise since it concerns itself with fields comprised $(\mathbb{Z}/p\mathbb{Z}, +, \times)$ Upon rereading, I see I jump into fairly complicated math fairly quickly.]

Group theory is the study of symmetry. Every set of symmetries on an object correspond to a group, and likewise every group corresponds to a set of symmetries.[0] Now to unpack what this means, it might be helpful to depict precisely what 'symmetry' is. The prototypical example is geometric symmetry. If we consider the set of symmetries of a square under rotations and reflections, we can form a group with a fancy name, the 'Dihedral group of order 8,' denoted D_8. (Order, here, refers to the number of elements of the group, or the cardinality/size of the underlying set.)[0] Group theory allows us to categorize the 'extent' of these symmetries as well. For instance, the group of all symmetries on a circle is, intuitively, much larger than that of the rectangle. In fact, this group is of infinite order (an 'infinite group')—much larger than the group of the symmetries of the square. This circle group, denoted $\mathbb{T}$, is an especially important one in many respects, for instance it

[0]Although this isn't a true "public-key" algorithm, DH can be easily modified to send arbitrary messages in a scheme known as 'Elgamal encryption.'
(**elgamal**)

[0]This is rigorously true if we consider the set of all automorphisms of an object (viz. the set of all bijective maps from and to itself), this forms a group under function composition with the inverses being the inverse maps and the identity being the identity "do nothing" map, so this perspective is justified.

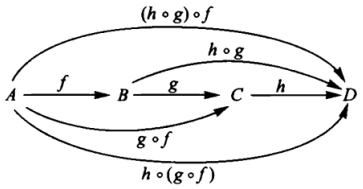
[0]The 'order' of an element also refers to the order/size of the subgroup generated by that element. Also, a group's order is allowed to be infinite.

establishes a certain duality between time and frequency, it behaves as a topological space, and it constitutes a major part of the Standard Model of Physics.

Now this interpretation of 'symmetry' is biased toward a very geometric perspective. If we want to broaden the horizons of group theory, we need to consider symmetries on objects like a law or principle, a mathematical equation, a structural rule, etc. In this interpretation, symmetry becomes a much deeper concept. For instance, our circle group above, $\mathbb{T}$, can also be represented as the set of all complex numbers with magnitude 1 under multiplication.

$$\mathbb{T} = (\{z \in \mathbb{C} \mid |z| = 1\}, \times)$$

Before I outright state the conditions that an object fits to be a group (the text-book definition), it might be helpful to see this type of thinking in action in the real-world. If we consider, as before, the 'law' or 'rule' that *justice is impartial*, what we're essentially saying is that "the verdict of a court case is independent of the qualities of the people involved, as in, permuting the names or qualities of the people involved has no effect on the outcome." Mathematically, this would correspond to invariance under a certain group action, where the group action is the one that does the permuting. This perspective of indistinguishably does hint at Group theory's widespread use in cryptography. In essence, we're saying "one cannot *a priori* give a property that holds for one instance and not another." To see where the textbook definition of a Group arises, we must first restrict the kinds of 'symmetries' we are looking for. Given a thing to be invariant and the type of transformations we're allowing, we can derive the group axioms like so: If we have two transformations T_1, T_2 that leave our object invariant we can see that their composition $T_1 \circ T_2$ must also leave the object invariant (composition referring to applying T_2 then applying T_1). Under this composition of transformations, one key property arises: *associativity*. It might be hard to see at first glance, but function composition is always associative: the order in which you evaluate functions has no effect on the result. This means we can drop parenthesis without any ambiguity, since $T_1 \circ (T_2 \circ T_3) = (T_1 \circ T_2) \circ T_3 = T_1 \circ T_2 \circ T_2$. This can be seen diagrammatically, for maps f, g, h below:



Beyond associativity, another requirement is the "do nothing" map. At

first seeming quite useless, it is the keystone of any group—the fundamental symmetry, if you will. One might interpret this as saying, "everything has a symmetry with itself." This element is called the *identity element* or the *identity map*. The last criterion for a group is that of inverses. In a sense, this requirement that every map has an inverse map boils down to an interpretation that, in order to be a 'true' symmetry, it must be invertible. This should make sense, since, for example, if we had a map that 'forgot' some properties of an object, there wouldn't be an inverse map[0] because this 'forgetful map' wouldn't be a bijection (one-to-one and onto). With that in mind, if one *doesn't* restrict themselves to invertibility, we get another algebraic structure called a 'Monoid.' But the study of these are much more unwieldy and untame than that of groups, so this is a reasonable restriction.

3.1 The Definition of a Group

A group is a set G, equipped with a binary operation mapping two elements to another of the form $* : \ G * G \to G$ such that the following conditions hold:[8]

Associativity
For all $a, b, c \in G, \ a * (b * c) = (a * b) * c$
Identity
There exists an $e \in G$ such that $a * e = e * a = a \ \ \text{for all } a \in G$
Inverse
Each $x \in G$ has a unique inverse $x^{-1} \in G$ with $x * x^{-1} = x^{-1} * x = e$ [0]

In our 'symmetry-focused' definition, this set consisted of the symmetry-preserving transformations on an object where the binary operation was composing transformations. It's worth mentioning that this axiomatic definition of a group was agreed upon some 100 years after mathematicians began to study groups. It is for this reason that the 'symmetry' understanding of a group can often be more applicable to answer "why study groups at all?" Under this axiomatic definition of a group, you could probably tease out a few examples. For instance, the integers form a group under addition $(\mathbb{Z}, +)$. It's worth verifying for yourself that these do fit the definition of a group. Our identity element is 0, since $0 + a = a + 0 = a$ for all $a \in \mathbb{Z}$. Addition of integers is clearly associative. And, every number has a unique inverse, $a^{-1} := -a$. This group can also be thought of as the sliding

[0] unless we expand our horizons to a notion of left-adjunction... :)

[8] **saracino**.

[0] Note that uniqueness of inverses is usually not given in the axioms for a group, since it follows from the Identity and Associativity requirement and the existence of a (not strictly unique) inverse. If b, c are left and right inverses respectivly of an element a with identity 1, then considering $c = 1 * c = (b * a) * c = b * (a * c) = b * 1 = b$, giving us $b = c$ as required. This technique also introduces a notion of a cancelative property to groups as well.

symmetries of the integer lattice on the number line, in which case I like to think of an operation $3 + 5$ corresponding to taking the number 5 and seeing where it ends up after sliding the entire numberline ahead 3 units (or vice-versa).[0] Similarly, the real numbers (excluding 0) form a group under multiplication. Again its worth verifying that it has an identity element 1, that it's operation is associative, and that every element has an inverse $a^{-1} := \frac{1}{a}$, since $a \times \frac{1}{a} = \frac{1}{a} \times a = 1, \ \forall a \in \mathbb{R}$. To help your intuition, also consider that the positive integers $(\mathbb{Z}, \times)$ *do not* form a group under multiplication, since inverses are not suitably defined, i.e., there is no integer a such that $a \times 2 = 2 \times a = 1$. The reason why the axiomatic approach is the predominant way to introduce groups is because by forgoing the need to study a group's representation, Group theory can concern itself more with the group itself and its properties in *any* situation it appears. Because of this, Mathematicians concern themselves with classifying groups *up to isomorphism*, meaning up to a renaming of the elements.[0]

3.2 The Case for Commutativity and Abelian Groups

One key property that we didn't require in our group axioms is *commutativity*. At first glance, this might seem like an obvious condition, since many of the introductory examples do obey commutativity. But most of the insight that comes from group theory is from the study of groups that do not obey commutativity. This distinction happens to be so important that we give groups obeying commutativity a special name: *Abelian groups*.[9] Likewise, groups that violate this condition are sometimes called *non-abelian groups*. In general, non-abelian groups can be thought of as corresponding to symmetries that change the backdrop for another symmetry to occur (obviously without violating associativity). For instance, the group of symmetries on a square, D_8, is non-abelian (meaning non-commutative). If you labeled the verticies of a square, you'd find performing a rotation and than a reflection is not, in general, the same as performing a reflection than a rotation[0]

[0]This intuition also highlights the generally non-commutative nature of groups. It's a remarkable fact that performing a specified slide on an integer is equivalant to performing a slide cooresponding to the value of the integer on the integer cooresponding to a slide, or (said in a different way), the fact that the set of integers is in a bijection with the set of all sliding transformations on a numberline—establishing a sort of duality between the two. In one case, the integer acts as an object and, in the other case, a transformation.

[0]It turns out if we restrict ourselves to finite *simple* groups (somewhat like the 'prime' building-blocks of groups), Mathematicians have already completed the classification. This was one of the major accomplishments in 20th century mathematics, provoking many questions about paradoxical 'sporadic groups.'
(**aschbacher**)

[9]**dummit**.

[0]If we were working with the presentation of D_8, $\langle a, x \mid a^4 = x^2 = e, \ xax^{-1} = a^{-1} \rangle$, we would say that x and a don't commute—corresponding to the reflection and rotation generators respectively.

—corresponding to the fact that reflections change the backdrop on which a rotation acts. Despite this, many of the other examples we've looked at so far happen to be abelian groups. For instance, the additive groups of the integers $(\mathbb{Z}, +)$, rationals $(\mathbb{Q}, +)$, and reals $(\mathbb{R}, +)$, all obey commutativity and are hence 'Abelian.' Similarly, the circle group $\mathbb{T}$, corresponding to the rotative symmetries of a circle, is commutative as well (a fact which is of great importance to Quantum Field Theory). Much of the groups that are of importance to cryptography are abelian groups. For intuition on why this may be the case, consider the RSA example before. Both parties combined a series of steps involving their public and private keys. If the order in which the modular exponentiation was performed mattered, Alice and Bob would get different results after combining their public and private keys in such a way.

4 Group Law on an Elliptic Curve

4.1 Definition of an Elliptic Curve

While the elliptic curve group admits an intuitive geometric description, it's important that our algebraic description satisfies the group axioms. $E(\mathbb{F}_q)$

In a complete turn of events (that will hopefully make sense in the end), I want to explain the notion of *Elliptic Curves*. Elliptic curves are a family of algebraic[10] curves defined as the solution set of a polynomial equation. Namely, an elliptic curve is the set of solutions (x, y) to an equation of the form:[11] [12]

$$y^2 = x^3 + ax + b, \text{ for constants } a, b.$$

Importantly, the behavior of this curve is dependent on the ground 'field' on which one defines it. We will look at the familiar case of the reals shortly, but this is an important distinction, hinting towards the possibility of a discrete context, or field, to come. Field, in mathematics, refers to a structure where any two elements have a well-defined notion of addition, subtraction, multiplication and division (except by a 0 element). The most important examples are the rationals $\mathbb{Q}$, the reals $\mathbb{R}$, and the complex numbers $\mathbb{C}$. Not the integers since, take, for example: $1, 2 \in \mathbb{Z}$, but $\frac{1}{2} \notin \mathbb{Z}$. We will begin with the case of an elliptic curve over the reals.

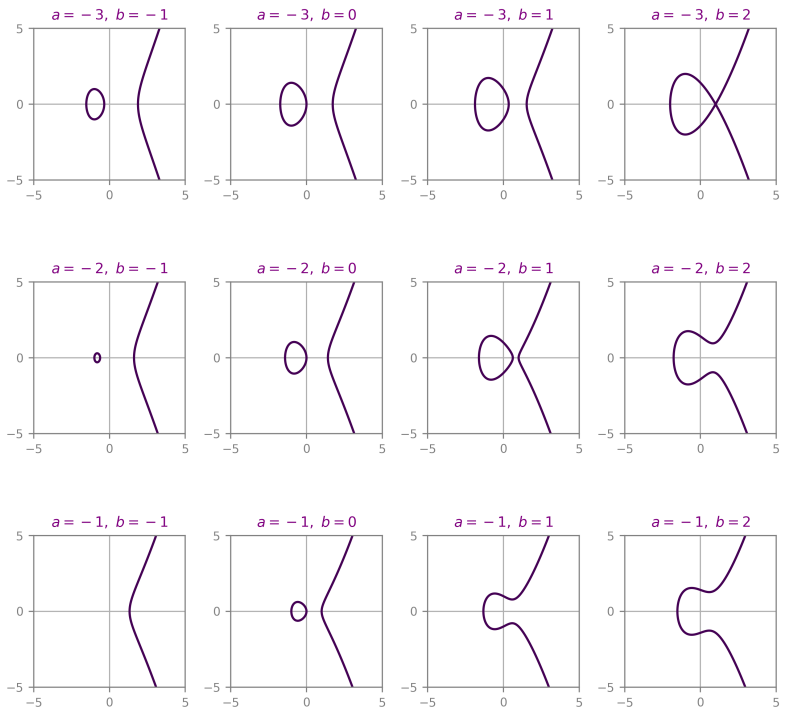


Figure 2: A few selected elliptic curves plotted with matplotlib over the reals with a from -3 to 0 and b from -1 to 2.

4.2 Elliptic Curves over the Reals

In 2, we see the general shape of an elliptic curve, usually with a sort-of bulbous head on the left. This complicated shape arises from square rooting a general depressed cubic of the form $y = x^3 + px + q$. Without the x^2 term, this cubic is guaranteed to have only one inflection point at $x = 0$, since its second derivative $y'' = 6x$ does not depend on p or q. Since square roots preserve the inflection points, we can see the connected graphs in 2 all

[10] Specifically a dimension one algebraic variety since it admits a group structure.

[11] **koblitz**.

[12] The general form of an elliptic curve $y^3 + a_1xy + a_2y = x^3 + a_3x^2 + a_4x + a_5$ can always be transformed in a series of scalings and rotations to the reduced form preserving the group structure as long as the field is of characteristic not 2 or 3.

have inflections at zero. Furthermore, the process of square-rooting increases values from $0 < y < 1$ and decreases values when $y > 1$ with a fixed point at 1. Over $\mathbb{R}$, square roots are limited to inputs greater than 0, but since we take both branches ($\pm$) of the root, we see reflection symmetry across the y-axis (this will be important in a moment). To illustrate the relationship:

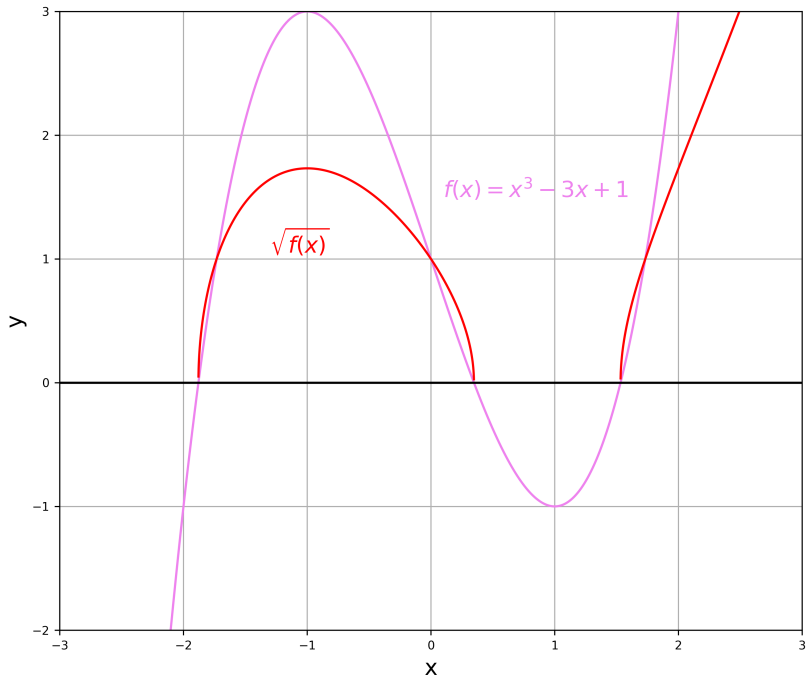


Figure 3: A depiction of how one branch of an elliptic curve arises from square-rooting a depressed cubic.

4.3 Geometric secant-tangent construction

It turns out, Elliptic curves are special because one can define a reasonable notion of addition with points on the curve. The one caveat is that we require a point 'at infinity' I to do so. Formally, this means were working in the projective real plane $\mathbb{RP}^2$ or $\mathbb{R}^2 \cup \{\infty\}$. But with this, we obtain

the property that any secant drawn through two points on the curve must intersect another third point on the curve or this point at infinity I. Thus we will define the addition of two points P and Q to be this third point reflected across the y-axis.[13]

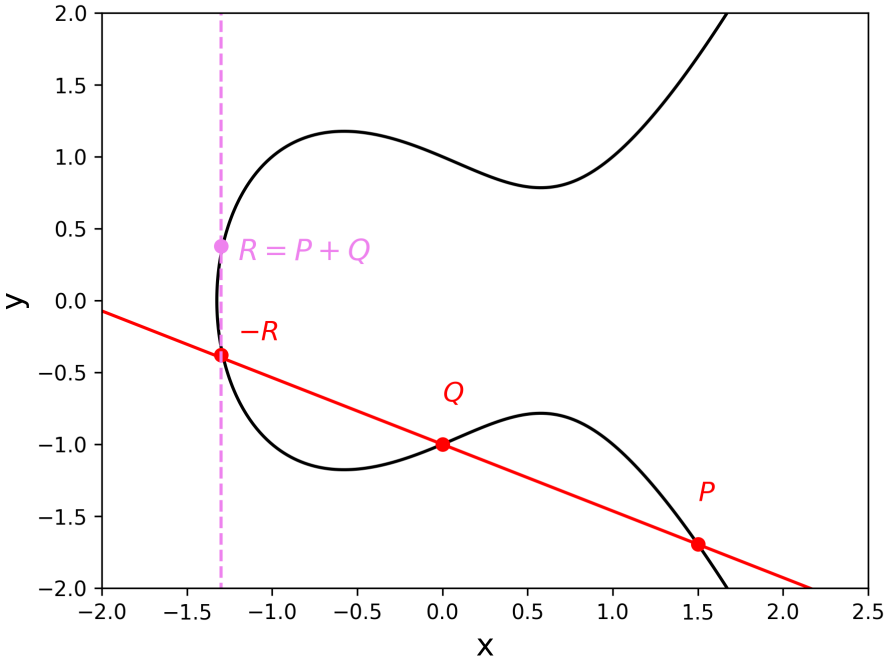


Figure 4: Point addition of two points P and Q, defined by reflecting the third point of intersection across the y-axis, on $y^2 = x^3 - x + 1$.

4.4 Addition of two points

We can make this idea rigorous by using some geometric tools. Our third point $-R$ is defined as the intersection of our elliptic curve $y^2 = x^3 + ax + b$ and some secant line $y = mx + l$. Given the points $P, Q = (x_p, y_p), (x_q, y_q)$ we can write down the slope:

$$m = \frac{y_p - y_q}{x_p - x_q}$$

[13]Note, without reflecting across the y-axis, this operation would not be associative. It also makes sense with the notion that three colinear points sum to the point-at-infinity $A + B + C = I$ and hence, $A + B = -C$.

Proceeding with the simultaneous equations

$$y^2 = x^3 + ax + b \tag{1}$$
$$y = mx + l \tag{2}$$

Substituting (2) into (1),

$$\begin{aligned}(mx+l)^2 &= x^3 + ax + b \\ m^2x^2 + 2mlx + l^2 &= x^3 + ax + b \\ \Rightarrow 0 &= x^3 - (m^2)x - (2ml)x + (b - l^2)\end{aligned} \tag{3}$$

Now, this cubic has three real solutions, namely the roots of P, Q and R respectively. Writing these as factors:

$$(x - x_p)(x - x_q)(x - x_r) = x^3 - (m^2)x^2 - (2ml)x + (b - l^2)$$

The next step is to expand the left side out and equate coefficients.

$$\begin{aligned}LH &= x^3 - (x_p + x_q + x_r)x^2 + (x_px_q + x_px_r + x_qx_r)x - x_px_qx_r \\ RH &= x^3 \qquad\quad - (m^2)x^2 \qquad\qquad\quad - (2ml)x + (b - l^2).\end{aligned}$$

This means, using the purple coefficients,

$$\begin{aligned}m^2 &= x_p + x_q + x_r \\ \therefore \quad x_r &= m^2 - x_p - x_q.\end{aligned} \tag{4}$$

Substituting the x coordinate of $-R$, x_r into the point-slope form of the line with $m = \frac{y_p - y_q}{x_p - x_q}$ we obtain,

$$\begin{aligned}y_r &= y_q + m(x_r - x_q) \\ &= y_p + m(x_r - x_p).\end{aligned}$$

Flipping our result across the y-axis to obtain R, we have (finally),[14]

$$\begin{aligned}P \;+\; Q \;&:=\; R \\ (x_p, y_p) + (x_q, y_q) &:= \left(m^2 - x_p - x_q, \quad -y_q - m(x_r - x_q)\right),\end{aligned}$$

$$\text{where } m = \tfrac{y_p - y_q}{x_p - x_q}.$$

■

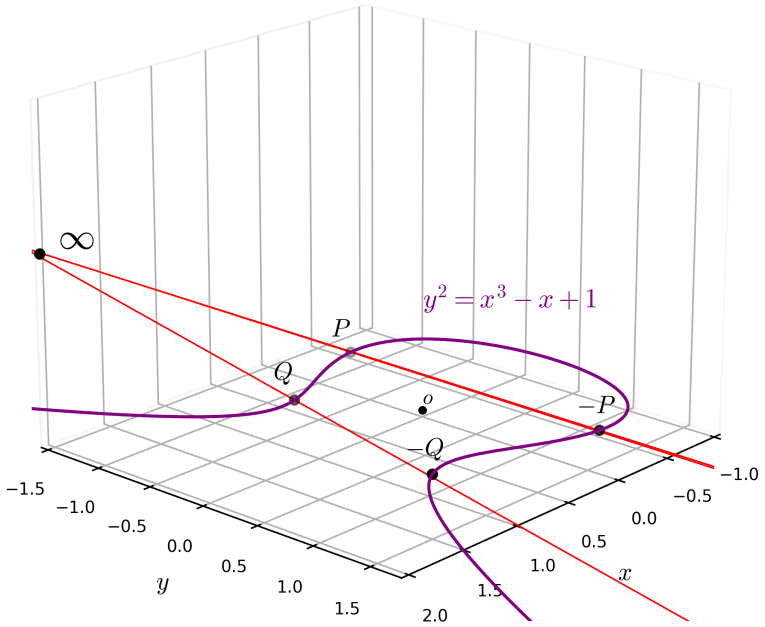


Figure 5: Parallel lines drawn through inverse points, all intersecting at point at infinity I over $y^2 = x^3 - x + 1$. As in, $P + (-P) = I = Q + (-Q)$.

4.5 Inverse Points and Infinity

In the case of adding two points opposite each other over the y-axis, one says they intersect the third point "at infinity." In general, over a projective plane, we revise the parallel postulate to say "any two parallel lines share one point of intersection, namely the point at infinity." We see this below; like train tracks meeting the horizon, the two parallel, vertical lines meet at ∞ or I.

Because of this construction, I acts as a left- and right-sided identity element. Since, for any ($\mathbb{R}$-rational) point A with inverse $-A$ on the elliptic-curve,

$$\begin{aligned} A + (-A) &= I \\ A + (-A) + A &= I + A \\ \Rightarrow \quad A + I &= I + A = A \end{aligned} \tag{5}$$

[14] As long as $P \neq Q$ and $P \neq -Q$.

4.6 Adding a point to itself

The case that was left out of the earlier pictorial representation of point addition is when a point is added to itself. Not to fear, however, since we can replace the secant through two points with a tangent through one. The resultant point is the resultant intersection flipped across the y-axis. The slope of this tangent can be found through implicit differentiation.

$$\begin{aligned} d(y^2) &= d(x^3 + ax + b) \\ 2ydy &= 3x^2dx + adx \\ \frac{dy}{dx} &= \frac{3x^2 + a}{2y}. \end{aligned}$$

Thus with,

$$y_r - y_p = m(x_r - x_p)$$

$$\text{where } m = \tfrac{3x_p^2+a}{2y_p},$$

we can use the same result as before with $P = Q$:

$$\begin{aligned} m^2 &= x_p + x_q + x_r \\ \Rightarrow \quad x_r &= m^2 - 2x_p \end{aligned} \tag{6}$$

and, using the point-slope form and flipping across the y-axis,

$$y_r = -y_p - m(x_r - x_p)$$

■

*A word on notation: the choice to use + to denote this binary operation is entirely arbitrary, and many authors use alternate notation to represent this ($\times, \oplus, \otimes, \star$ etc.). Using + hints that adding a point P to itself n times might be written nP, but for sake of consistency with the theory to come, I will write this P^n meaning $P + P + ... + P$, n-times. This will make sense in a moment.

5 Group structure of E(K)

5.1 Verifying the group axioms

By the higher powers of mathematical serendipity, this forms a group! A very special group at that. We can write this elliptic curve group as $E[\mathbb{R}]$ for the case in the reals. Or, $E[K]$ for any field K (again, a structure with $+, -, \times, \div$).

Now, we will (at least, heuristically) verify that the group axioms hold, we first see in eq (5) that we have a two-sided identity, I, given by the point at infinity. Hence,

$$P + I = I + P = I \text{ for all } P \in E[K]$$

Next, we prove the existence of an inverse for every element by noticing that our curve is symmetric about the y-axis. Similarly our identity element I is self-inverse. Hence,

$$\text{for all } P \in E[K], \text{ there exists an inverse } -P \text{ such that } P + (-P) = I$$

This operation is associative as well, but unfortunately a rigorous proof of this fact is unprecedentedly untame, especially through this geometric depiction of elliptic curves. With more powerful tools from algebraic geometry, this proof is almost trivial, following nicely from tracking values on points of a curve (namely $\text{Div}^0(E)$) and the relationship between poles and zeros, as stated by the Riemann-Roch theorem.[15] Lastly, this binary operation of adding two points is *commutative* as well, following geometrically from the secant construction or through the formulas.

As a result of meeting these conditions, $E[K]$ forms an abelian group under our binary operation.

5.2 Elliptic curves over finite fields, $K = \mathbb{F}_p$

This elliptic curve group action holds for other fields K as well. Most important to cryptography is elliptic curves over "finite" fields. A finite field, denoted $\mathbb{F}_p$ for p prime[16], is a set of ordered pairs (x, y) with addition, subtraction and multiplication taken *modulo* p. Division takes more care to define and is, in fact, the reason why $\text{char}(E)$ must be prime.[17] Multiplicative inverses (division) can be computed using the Euclidean Algorithm or by leveraging the symmetry. Essentially, over a field $\mathbb{F}_p$, the multiplicative inverse of a number n is a number m such that,

$$nm \equiv 1 \mod p$$

The invertible elements of $\mathbb{F}_p$ are exactly those that are relatively prime to p. The proof is by contradiction:

If $\gcd(n, p) = d$ with $d \neq 1$, assume there is some m for which $nm \equiv 1$

[15] **silverman**.

[16] Technically, p can be a prime power. One might write $\mathbb{F}_q$ where $q = p^r$.

[17] **saracino**.

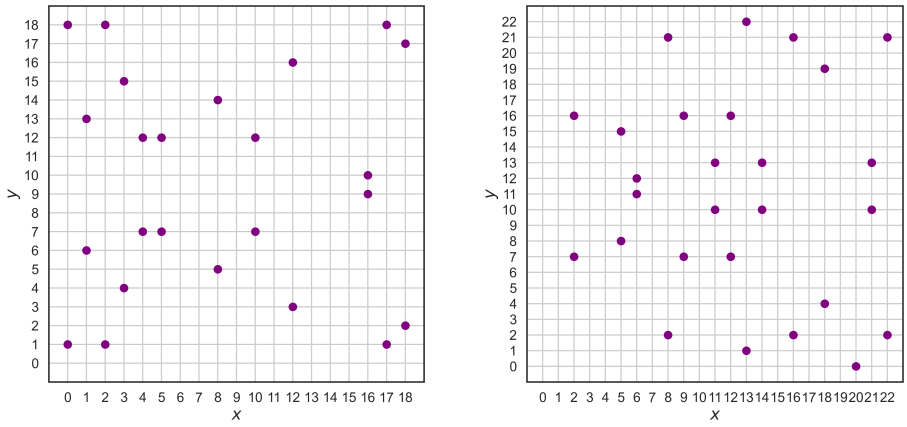
mod p. Expanding this out,

$$\begin{aligned} nm &\equiv 1 \mod p \\ nm - 1 &\equiv 0 \mod p \\ \Rightarrow\ & p|nm-1 \\ \Rightarrow\ & d|nm-1 \\ \Rightarrow\ & d|nm-1-nm \text{ (since } d|nm) \\ \Rightarrow\ & d|1 \end{aligned}$$

As 1 is not divisible by anything except 1, our assumption was false and $\gcd(n,p)$ must be one.[18] ■

With this in mind, one can compute the elements/points of $E[\mathbb{F}_p]$ as,

$$E[\mathbb{F}_p] = \left\{(x,y) \mid y^2 \equiv x^3 + ax + b \mod p\right\} \cup \{\infty\}.$$



(a) $y^2 = x^3 + 15x + 1$ over $\mathbb{F}_{19}$ (b) $y^2 = x^3 + 12x + 17$ over $\mathbb{F}_{23}$

Figure 6: Select elliptic curves plotted over finite fields. Notice the symmetry about $\frac{p}{2}$.

6 Conclusion

6.1 Limitations and Reflection

[18] **koblitz**.