Crypto Theoretical Minimum
Elliptic Curve Groups

Bassam El Khoury Seguias

BTC: 3FcVvBZwTUkUrcqJd16RcjR42qT2tDWHWn

ETH: 0xb79Fb9194C8Cc6221368bb70976e18609Ab9AcA8

May 24, 2018

1 Introduction and motivation

The sempiternel question of how to gain and maintain power has haunted the minds of humanity's brightest and darkest since the dawn of civilization. Be it physical (e.g., military) or economical (e.g., wealth), power's very existence relied in part on access to information. Asymmetric information that is. Numerous are history's examples that demonstrate how entities that knew what others didn't and that were able to act on it, benefited from an unfair advantage. The quest for sustainable power motivates the protection of one's proprietary information and the attempt at breaching that of the others.

Although significant in its own right, the pursuit of power is not the only motivator to conceal information. Privacy, in so far as the individual's well-being is concerned, is another. In that respect, two areas stand out. The first is concerned with the unique nature of a human persona. As a matter of observation, and at the risk of irritating adherents of monism, the attributes of a human personality are so varied. Each attribute exists on a wide spectrum, making it unlikely that any two individuals have the same profile so to speak. The privacy spectrum is no exception, and while some live their lives as an open book, others might not even be comfortable sharing their half title page. The second area is concerned with the safety of a certain subset of individuals, e.g., whistle-blowers. They may hold sensitive information destined to be shared with a specific party. Should this information fall in the wrong hands, it could jeopardize the safety of the source.

It is therefore reasonable to assume that not every piece of information is meant to be common knowledge. One could certainly debate the merits of such a claim and in the process, revisit the very foundation of power, privacy and safety. The fact remains however, that information can be a source of influence, discomfort, and danger. One way of protecting specific content and limiting its access to intended parties only, is through the use of encryption and decryption algorithms.

Symmetric-key vs. public key cryptography Encryption can be thought of as a map that takes a relevant piece of data known as a message, and outputs an altered version of it. The map can be either one of two types: 1) **PT-invertible**, or 2) **One-way**.

1. We say that a function is PT-invertible if one can find a polynomial-time algorithm to compute its inverse. In order for a sender to encrypt a message, he must be able to compute the map. On the other end, in order for a recipient to decrypt the message, she must be able to compute the inverse map. If a map is PT- invertible, information used to build it is sufficient to build its inverse. As a result, senders and recipients alike can use the same information to encrypt and to decrypt messages in a process known as **symmetric-key cryptography**. The symmetric information shared between the sender and the recipient is known as the **secret key**.

   As an example, consider a message space consisting of the case-agnostic latin alphabet of 26 letters. We represent each letter by its numerical equivalent (e.g., letter "a" or "A" represented by 1), and apply the following affine map:

$$f : \{1,\ ...,\ 26\}^* \to \{1,\ ...,\ 26\}^*$$

$$(n_1, n_2, ...) \to f(n_1, n_2, ...) \equiv (\alpha n_1 + \beta \pmod{26},\ \alpha n_2 + \beta \pmod{26},\ ...)$$

   where the $*$ superscript denotes a string of arbitrary length, and $\alpha,\ \beta$ are pre-defined elements in $\{1,\ ...,\ 26\}$ such that α is relatively prime to 26.

   For instance, let $\alpha = 3$ and $\beta = 5$. The word "chaos" has a representation given by (3, 8, 1, 15, 19). When fed to the map, one obtains an output given by $f(3, 8, 1, 15, 19) = (14, 3, 8, 24, 10)$. This corresponds to "nchxj".

   The inverse map can be written as follows:

$$f^{-1} : \{1,\ ...,\ 26\}^* \to \{1,\ ...,\ 26\}^*$$

$$(y_1, y_2, ...) \to f^{-1}(y_1, y_2, ...) = (\alpha^{-1}(y_1 - \beta) \pmod{26},\ \alpha^{-1}(y_2 - \beta) \pmod{26},\ ...)$$

   where α^{-1} is the inverse of α in modulo 26 arithmetic (i.e., $\alpha \times \alpha^{-1} \equiv 1 \pmod{26}$). Since, α and 26 are relatively prime, one can use the extended euclidean algorithm outlined in the *Groups and Finite Fields* post to calculate α^{-1} in polynomial time. Consequently, f is a PT-invertible map. All that is required to build it and its inverse is the pair (α, β). This pair constitutes the symmetric key shared between the sender and the recipient.

   Symmetric-key cryptography is known to be efficient, with the possibility of encrypting and decrypting large amounts of data relatively fast. Its weakness however, lies in the fact that the secret key must be shared between two (or more) parties over a **secure channel**. Enforcing perfect security and eliminating the risk of leakage over a digital communication channel is a challenging endeavour. Moreover, if such a secure medium of communication could be constructed, it would be legitimate to question the usefulness of sharing a secret key in the first

place as opposed to using the secure channel to directly send and receive the actual messages.

2. One-way or trapdoor functions have no known polynomial-time algorithm to compute their inverses in the absence of a specific piece of information referred to as the **trapdoor**. In other words, knowledge of the building blocks of a map is not sufficient to compute its inverse. In order to do so (and as a result decrypt a message), a recipient must have access to the trapdoor.

Algorithms where the information needed to encrypt is different than that needed to decrypt, form the basis of **asymmetric cryptography**. The nomenclature is a reflection of the informational asymmetry between encryption and decryption. More specifically, each recipient is associated with a key pair consisting of a unique **private key** only known to her, and a related **public key** that can be shared with anyone. Anyone can use the public key of a recipient to encrypt a message. Decryption however, requires knowledge of the private key which is only known to the recipient. In light of the above, a crucial criterion in the design of key pairs is that no entity should be able to derive the private key from the public one. The dual-key architecture is the reason why asymmetric cryptography is also known as **public-key cryptography**.

As an example we consider the RSA encrytion scheme. RSA generates the public and private keys of a user as follows:

- Select two very large primes p and q such that $p \neq q$.
- Let $n \equiv p \times q$. One can observe that given p and q, it is easy to compute n. However, given n, it is extremely challenging to find p and q. This is known as the factoring problem, thought to be intractable on the group $((\mathbb{R}^+)^*, \times)$.
- Find n's totient value $\phi(n)$. Euler's totient function returns the number of integers less than or equal to n that are relatively prime to n. If n is prime, then $\phi(n) = n - 1$. In addition, for any two coprime numbers p and q, $\phi(p \times q) = \phi(p) \times \phi(q)$. Consequently,

$$\phi(n) \;=\; \phi(p \times q) \;=\; \phi(p) \times \phi(q) \;=\; (p-1) \times (q-1)$$

- Choose $e \in \mathbb{N}^*$ such that $1 < e < \phi(n)$ and such that $\gcd(e, \phi(n)) = 1$. We let (n, e) be the public key of the user.
- Choose $d \in \mathbb{N}^*$ such that $d \times e \equiv 1 \pmod{\phi(n)}$. That means that $d \times e = 1 + \alpha \times \phi(n)$ for some integer α. We observe that even if e were known, it is computationally hard to calculate $\phi(n)$ because this requires computing $(p-1)$ and $(q-1)$ which in turn, requires solving the prime factorization problem. This in turn, makes it challenging to calculate d. We let (n, d) be the private key of the user. The previous observation shows that it is unlikely for anyone to derive the private key (n, d) from the public key (n, e) alone.

To encrypt a message destined to Bob, Alice first transforms it into an integer $0 < m < n$ by using a common-knowledge pre-defined mapping. We require that m and n be coprime. Subsequently, Alice computes the encrypted value $v = m^e \pmod n$, where (n, e) is Bob's public key. In order to decrypt the message, Bob uses his private key (n, d) to compute $v^d \pmod n$. To see why this works, note the following equalities:

$$v^d \pmod n \;=\; m^{ed} \pmod n \;=\; m^{1+\alpha\phi(n)} \pmod n \;=\; m \times (m^{\phi(n)})^\alpha \pmod n$$

We can invoke Euler's theorem that states that if m and n are relatively prime, then $m^{\phi(n)} \equiv 1 \pmod n$ (proof ommitted). We conclude that

$$v^d \pmod n \;=\; m \pmod n$$

Note that when the public key (n, e) is known, it is straightforward to compute $v = m^e \pmod n$. However, calculating its inverse (i.e., finding the value of m when v and (n, e) are known) is thought to be hard. On the other hand, knowledge of the private key (n, d) allows a quick retrieval of m as we saw above.

A downside of public key cryptography is that it is not nearly as efficient as its symmetric counterpart, especially as the message size increases. However, symmetric-key cryptography depends on the existence of a secure channel which is challenging to build. The upside of asymmetric cryptography is that it bypasses the need for a secure channel altogether. It turns out that one can leverage the advantage of each type of cryptography to create a hybrid system that is both secure and efficient. This is accomplished through the use of a key-exchange protocol known as a **Diffie-Hellman exchange**.

The idea is to simply apply public-key cryptography to communicate a shared secret, which can then be used in a symmetric-key setting to encrypt and decrypt larger messages. Since the secret-key is relatively small (from a data standpoint), it can be encrypted rather efficiently using a public-key setting and then shared with a recipient on an untrusted channel. Larger message blocks can subsequently be effectively encoded and decoded using the secret key. More formally, an example of this setting can be described as follows:

- Consider the multiplicative cyclic group $(\mathbb{Z}_p^* \,, \otimes)$ of the finite field $(\mathbb{Z}_p \,, \oplus, \otimes)$ of large prime order p (The reader can consult the *Group and Finite Fields* post for more details). Let g be one of its generators.
- Alice and Bob chose their individual secret keys $sk_a,\ sk_b$ from $\mathbb{F}_p^* \equiv \{1,\ ...,\ p-1\}$
- They compute their respective public keys

$$pk_a = g^{sk_a} \equiv g \otimes g \otimes ... \otimes g \ (sk_a \text{ times})$$

$$pk_b = g^{sk_b} \equiv g \otimes g \otimes ... \otimes g \ (sk_b \text{ times})$$

- Bob uses Alice's public key to compute $pk_a^{sk_b} = g^{sk_a \times sk_b}$. Similarly, Alice uses Bob's public key to compute $pk_b^{sk_a} = g^{sk_b \times sk_a}$.

- The two previously calculated values are equal and known only to Alice and Bob. This is because its calculation requires knowledge of at least one of the two secret keys. It can then be used as a shared secret as part of a symmetric-key algorithm.

As noted earlier, the most important design criterion is to ensure that the secret key cannot be derived from the public key. In our setting, this means that when given g and $pk = g^{sk}$, no one should be able to calculate in polynomial time the value of sk. This is known as the **discrete logarithm** (DL) problem and we will revisit it. On the multiplicative group of a finite field of large prime order, the DL problem is thought to be hard.

Digital signatures Encryption schemes help protect the content. However, they provide no proof that a certain sender was the actual author. This is true especially in the context of public-key cryptography where encryption keys are made public, allowing any party to claim that it was the actual sender. This problem can have drastic consequences when dealing with cryptocurrencies. Indeed, a cryptocurrency transaction consists of a message whose content allows a transfer of spending control from one owner to another. In Bitcoin for example, all valid transactions are publicly registered on the blockchain, and their content is purposefully not encrypted in order to enforce transparency and allow nodes to validate or reject them. However, the message in this case must be accompanied by a proof that the sender is actually the initiator of the transaction. Otherwise, anyone could initate a transaction on behalf of someone else without their consent, potentially causing financial mayhem.

The authentication process is done through the use of a mathematical construct known as a digital signature. In the context of cryptocurrencies, we care about digital signature and less so about encryption. The most important attribute of a digital signature is that of unforgeability. This can be defined in a variety of ways, but for all practical matters we mean resilience against **existential forgery in the adaptive chosen-message attack**. More details about digital signatures and the definitions of forgery can be found in the post entitled *Digital Signature and Other Prerequisites*. Generally speaking, digital signatures use the same public-key cryptography infrastructure described earlier for encryption and decryption. The sender signs with her private key in order to authenticate the message. Anyone on the network can then verify that she was the actual sender by running a verification algorithm that relies on the sender's public key. Various examples of digital signatures including Schnorr, RSA, generic Pointcheval & Stern models, as well as a number of more elaborate ring signature schemes can be found in previous posts.

The discrete logarithm problem A necessary condition to avoid forgery is that no one should be able to derive the private key from the public one. Here too, we realize the cruciality of one-way constructs. An important example of such a construct is the one associated with the DL poblem encountered earlier. The hardness of the DL problem on some well-defined groups underlies the security of various digital signature schemes, including those adopted in cryptocurrencies. Formally, we define the DL problem as follows:

- Let $(G, \otimes)$ be a given group and g an element of G.

- Given an element $y \in \{g\}$ (i.e., the subgroup generated by g), find an integer $x \in \mathbb{N}$ such that $y = g^x \equiv g \otimes g \otimes ... \otimes g$ (x times).

The smallest x that satisfies the above equation is known as the logarithm of y in base g and we write $x = log_g(y)$.

The difficulty associated with calculating x knowing y and g depends on the underlying group $(G, \otimes)$. On some groups, the problem is easy to solve (i.e., we know of polynomial-time algorithms that can solve it). On others, it is harder. Moreover, there exists different levels of difficulty, the highest being exponential (i.e., the only known algorithm(s) to solve the problem are exponential in time). In the context of public-key cryptography, it is always desirable to operate on groups where the hardness of the DL problem is exponential.

An example of a group on which the DL problem is easy to solve is $(\mathbb{Z}_n, \oplus)$ (i.e., the group of integers modulo n introduced in the *Group and Finite Fields* post). To see why, first note that this group is cyclic and that the equivalence class [1] is a generator. Given $[y] \in \mathbb{Z}_n$, the DL problem consists in finding $x \in \mathbb{N}$ such that

$$[y] = [1]^x \equiv [1] \oplus [1] \oplus ... \oplus [1] \ (x \text{ times})$$

By the definition of $\oplus$, this is equivalent to finding x such that $[y] = [x]$ i.e., y such that $y \equiv x \pmod{n}$. Consequently, one can compute x efficiently using the Euclidean algorithm.

An example of a group on which the DL problem is believed to be hard is the multiplicative group $(\mathbb{Z}_p^*, \otimes)$ of the finite-field $(\mathbb{Z}_p, \oplus, \otimes)$ of large prime order p. This group is cyclic and was introduced in the *Group and Finite Fields* post. The time required by the best-known algorithm to solve DL on it is $\sim \mathcal{O}(e^{c\sqrt[3]{(log\ p)(log\ log\ p)^2}})$ [10]. The running time is sub-exponential i.e., executes faster than an exponential algorithm but is less efficient than a polynomial one.

Despite the hardness of DL on the multiplicative cyclic subgroup of a large finite field, it remains more desirable to operate on groups where the DL is thought to be exponentially hard. An example of such a group is the one associated with **elliptic curves** over finite fields. The elliptic curve discrete logarithm problem (ECDLP) over $\mathbb{F}_p$ is thought to be exponentially hard, with the best performing algorithm requiring time $\sim \mathcal{O}(\sqrt{p})$ [10].

By way of comparison, ECDLP on a finite field of order $\sim 2^{160}$ has an equivalent difficulty to a DL problem on the multiplicative cyclic subgroup of $\mathbb{F}_{\sim 2^{1000}}$. One implication is that cryptographic primitives based on ECDLP require significantly smaller keys. This explains why the digital signature schemes used in various cryptocurrencies (e.g., Bitcoin's ECDSA, Monero's MLSAG) rely on elliptic curve groups.

With this motivation behind us, we are now in a position to introduce the concept of an elliptic curve. Its theory is very rich and sits at the intersection of different branches

of mathematics including analysis, geometry, and algebra. Our objective is to build a group structure based on the geometry of elliptic curves. The new group is referred to as an elliptic curve group and forms the public-key infrastructure of a number of cryptocurrencies in use today. We highlight that this introduction is limited to the minimum that we think is needed to appreciate the subject. It is by no means a comprehensive treatise. Readers interested in a detailed treatment of elliptic curve theory can consult e.g., [10]

In what follows, we first introduce an analytic view of elliptic curves over arbitrary fields. We describe the general Weierstrass form and derive a more simplified version as long as some constraints are observed. We then look at the geometry of elliptic curves over real numbers, and build a group structure after augmenting the curve with a point at infinity. The group's binary operation, also referred to as point addition, is described geometrically and analytically. Later, we introduce the elliptic curve group over finite fields and finally, we describe the two elliptic curves used in Bitcoin and in Monero.

2 Elliptic curves: An analytic description

One way of defining an elliptic curve is as a set of points satisfying the Weierstrass general equation and given by:

$$E : y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$$

The coefficients a_1, a_2, a_3, a_4, a_6 are chosen from a field $\mathbb{K}$ and we say that E is defined over $\mathbb{K}$. Note that a_5 is purposefully left out for reasons that we keep out of scope for now. $\mathbb{K}$ could be for instance the field of real numbers $\mathbb{R}$, or any finite field. Recall that in the *Groups and Finite Fields* post we mentioned that any finite field is either of prime order p or is an extension $\mathbb{F}_{p^m}$ of a field of prime order p where m can be any positive integer. We refer to p as the characteristic of the finite field $\mathbb{F}_{p^m}$ and write $char(\mathbb{F}_{p^m}) = p$. We derive below a simplified version of the Weierstrass equation applicable only if we exclude fields of characteristics 2 and 3.

Let's first look at the left-hand side of the equation. It is tempting to complete the quadratic in y. We can always find λ such that

$$(y + \lambda)^2 - \lambda^2 = y^2 + a_1xy + a_3y$$

We could subsequently make a change of variables by substituting y with $u \equiv y + \lambda$. Since λ does not depend on y, we would have eliminated all terms that contain y as a factor. The aforementioned equation in λ can equivalently be written as

$$2y\lambda = a_1xy + a_3y$$

One would then be tempted to conclude that $\lambda = \frac{a_1x+a_3}{2}$, except that division by 2 is not always permissible on an arbitrary field $\mathbb{K}$. If $char(\mathbb{K}) = 2$, then $2 \equiv 0 \pmod 2$ will not admit a multiplicative inverse on $\mathbb{K}$. However, division by 2 is possible on all other fields. In what follows, we always assume that $char(\mathbb{K}) \neq 2$. Consequently, we can

compute λ and perform the change of variable. The Weierstrass equation becomes:

$$u^2 - (\tfrac{a_1x+a_3}{2})^2 = x^3 + a_2x^2 + a_4x + a_6$$

$$\iff u^2 = (\tfrac{a_1x}{2})^2 + (\tfrac{a_3}{2})^2 + (\tfrac{a_1a_3x}{2}) + x^3 + a_2x^2 + a_4x + a_6$$

$$\iff u^2 = x^3 + (a_2 + \tfrac{a_1^2}{4})x^2 + (a_4 + \tfrac{a_1a_3}{2})x + (a_6 + \tfrac{a_3^2}{4})$$

Letting $a_2' \equiv a_2 + \frac{a_1^2}{4}$, $a_4' \equiv a_4 + \frac{a_1a_3}{2}$, and $a_6' \equiv a_6 + \frac{a_3^2}{4}$, the elliptic curve equation becomes:

$$E : u^2 = x^3 + a_2'x^2 + a_4'x + a_6'$$

The next step consists in simplifying the right hand-side of this equation. It turns out that any cubic equation can be transformed into an equivalent one with the quadratic term eliminated. We do so by substituting variable x with a variable of the form $v = x + \nu$. The value of ν is derived by first performing the substitution and then eliminating the coefficient of the quadratic term as follows:

$$u^2 = (v - \nu)^3 + a_2'(v - \nu)^2 + a_4'(v - \nu) + a_6'$$

$$\iff u^2 = v^3 - 3v^2\nu + 3v\nu^2 - \nu^3 + a_2'v^2 - 2a_2'v\nu + a_2'\nu^2 + a_4'v - a_4'\nu + a_6'$$

$$\iff u^2 = v^3 + (a_2' - 3\nu)v^2 + (3\nu^2 - 2a_2'\nu + a_4')v - \nu^3 + a_2'\nu^2 - a_4'\nu + a_6'$$

We require that the coefficient of v^2 be equal to 0. This imposes a constraint on ν's value which must satisfy:

$$-3\nu + a_2' = 0$$

If $char(K) \neq 3$, we can always find a multiplicative inverse of 3 in $\mathbb{K}$ and as a result, solve for $\nu = \frac{a_2'}{3}$. The elliptic curve equation becomes:

$$u^2 = v^3 + (-\tfrac{(a_2')^2}{3} + a_4')v + (a_6' + \tfrac{2(a_2')^3}{27} - \tfrac{a_2'a_4'}{3})$$

We can relabel the (v, u) variables as (x, y), and let $A = (-\frac{(a_2')^2}{3} + a_4')$ and $B = (a_6' + \frac{2(a_2')^3}{27} - \frac{a_2'a_4'}{3})$. We then obtain the simplified Weierstrass equation of an elliptic curve over a field $\mathbb{K}$ such that $char(\mathbb{K}) \notin \{2, 3\}$:

$$E : f(x, y) = y^2 - x^3 - Ax - B = 0$$

In the following section we construct a group structure over elliptic curves. The tangent to the curve at a given point will play an essential role in this construction. As a result, elliptic curves that have singularities (i.e., points where the curve is not differentiable) are not desired and will be excluded. Examples of singularities on a curve include cusps and self intersections. Analytically, a necessary and sufficient condition for a point $P \equiv (x_p, y_p)$ on a curve $f(x, y) = 0$ to be singular is for the partial derivatives at (x_p, y_p)

to be equal to 0. For the elliptic curve equation we get:

$$\{\ f(x_p, y_p) = 0 \iff y_p^2 - x_p^3 - Ax_p - B = 0$$

$$\{\ f_x(x_p, y_p) = 0 \iff -3x_p^2 - A = 0$$

$$\{\ f_y(x_p, y_p) = 0 \iff 2y_p = 0$$

The last equation implies that $y_p = 0$. If we substitute $y_p = 0$ in the first and second equations, we conclude that

$$P \equiv (x_p, y_p) \text{ is singular } \Rightarrow\ x_p^3 + Ax_p + B = 0 \text{ and } 3x_p^2 + A = 0$$

Consequently, x_p must be a cubic root of $x^3 + Ax + B$ as well as of its derivative $3x^2 + A$. This means that x_p is a double root of $x^3 + Ax + B$. If we let β denote the third root, we get the following factorization:

$$(x^3 + Ax + B) = (x - x_p)^2(x - \beta)$$

$$= (x^2 + x_p^2 - 2x_p x)(x - \beta) = x^3 - (2x_p + \beta)x^2 + (x_p^2 + 2x_p\beta)x - x_p^2\beta$$

By comparing coefficients, we find that $\beta = -2x_p$, $A = -3x_p^2$ and $B = 2x_p^3$. This in turn implies that the discriminant $\Delta \equiv 4A^3 + 27B^2 = 0$. To summarize, we showed that given an elliptic curve $E : f(x, y) = y^2 - x^3 - Ax - B = 0$ over a field $\mathbb{K}$ such that $char(\mathbb{K}) \notin \{2, 3\}$, we have:

$$P \equiv (x_p, y_p) \in E \text{ is singular } \Rightarrow\ \Delta \equiv 4A^3 + 27B^2 = 0$$

The contrapositive statement allows us to derive a sufficient condition for E to be non-singular. Specifically, if $\Delta \neq 0$ for all $(x, y) \in E$, then E is non-singular. Going forward, we only consider non-singular elliptic curves defined over fields of characteristic other than 2 or 3:

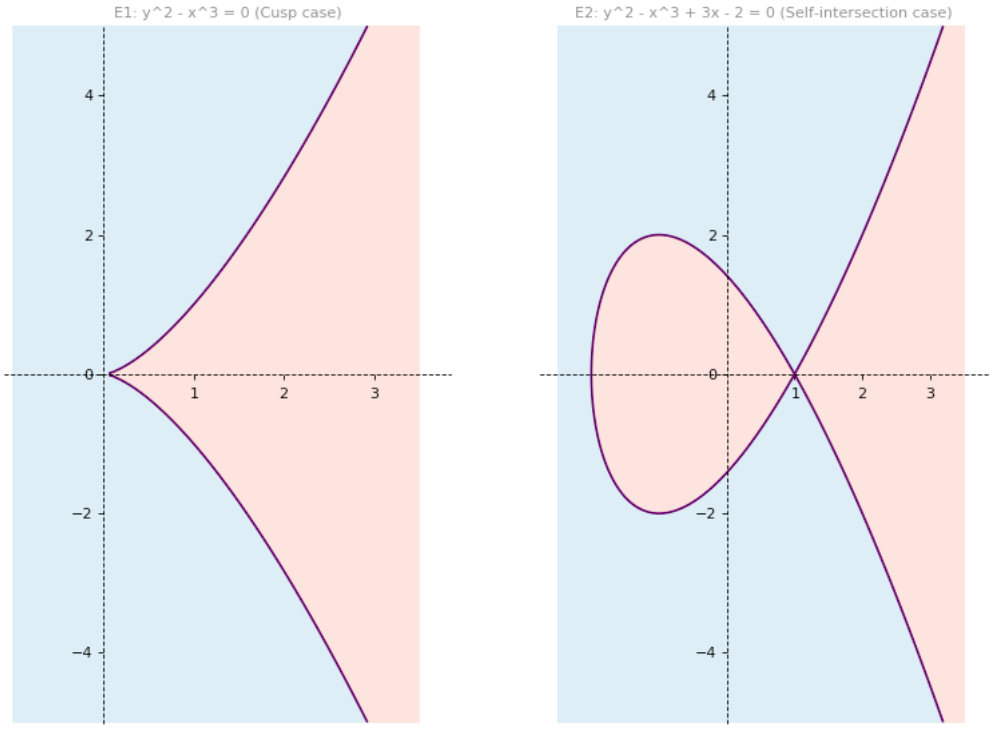
$$E = \{(x, y) \in \mathbb{K} \mid (y^2 = x^3 + Ax + B)\ \cap\ (char(\mathbb{K})\ \notin \{2, 3\})\ \cap\ (4A^3 + 27B^2 \neq 0)\}$$

3 Elliptic curve groups: A geometric approach

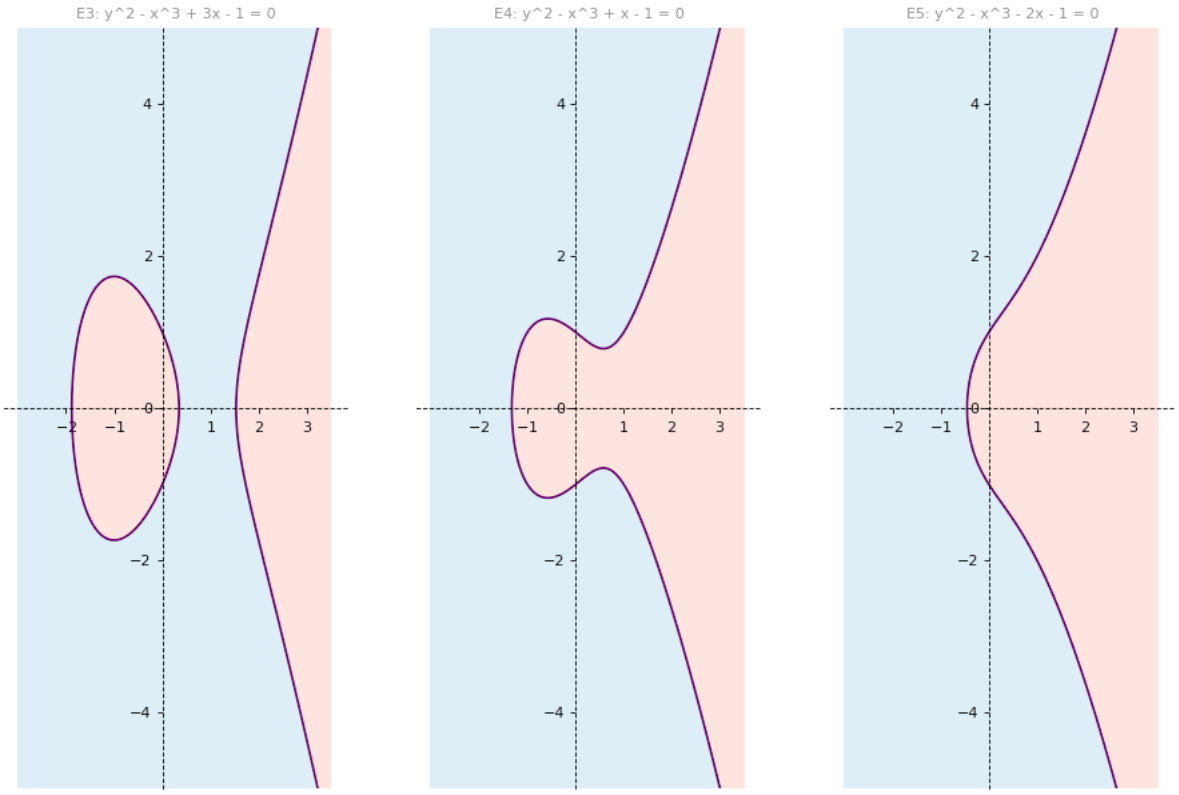
In what follows, we endeavour to build the elliptic curve group over finite fields. To do so, we first consider elliptic curves over $\mathbb{R}$ and study their geometry in order to devise a natural abelian group structure. Technically, the construction is performed in the projective plane as opposed to the euclidean plane. However, we attempt to motivate and justify the build-up without delving into the technicalities of projective geometry. Finally, we adapt the binary operation of the group to the case of a finite field.

Elliptic curves over $\mathbb{R}$ can be easily drawn on the euclidean plane. We include below the graphs of five different elliptic curves, two of which are singular and three regular.

SINGULAR ELLIPTIC CURVES



NON-SINGULAR ELLIPTIC CURVES



Elliptic curves exhibit x-axis symmetry. To see why, note that $\forall P \equiv (x_p, y_p)$ on the curve, it must hold that $\overline{P} \equiv (x_p, -y_p)$ is also on the curve. Indeed,

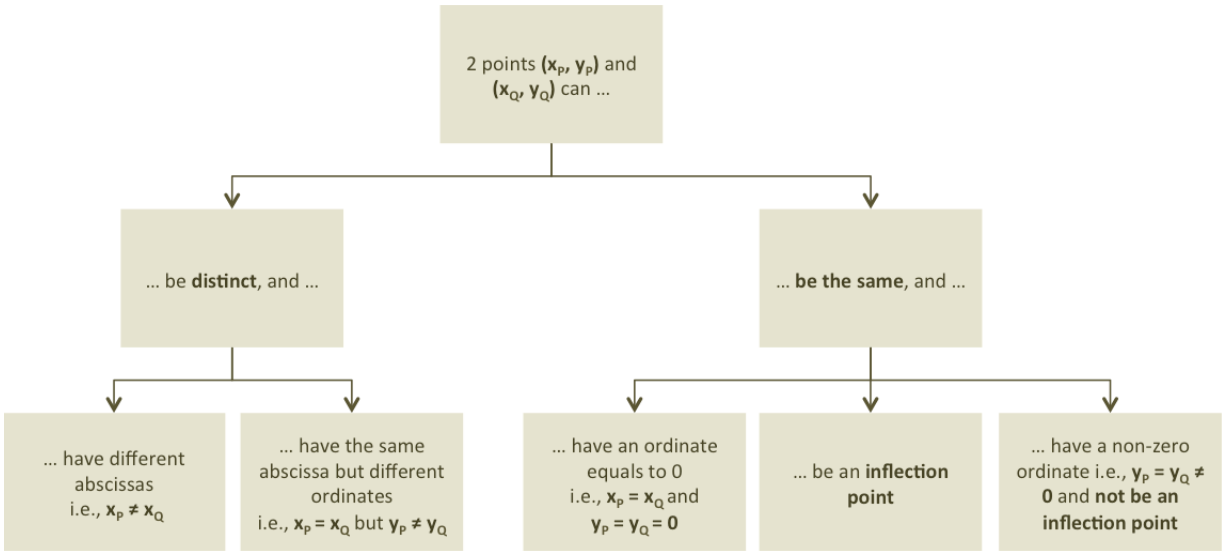
$$y_p^2 - x_p^3 - Ax_p - B = (-y_p)^2 - x_p^3 - Ax_p - B$$

Moreover, $y_p^2 - x_p^3 - Ax_p - B = 0$ by virtue of P being a point on the curve. Therefore, $(-y_p)^2 - x_p^3 - Ax_p - B = 0$, demonstrating that $\overline{P}$ is also a point on the curve.

In order to define any group, one needs to have an underlying set of elements as well as a binary operation on it that ensures that the group axioms are observed. In our case, the underlying set contains all the points in the euclidean plane that satisfy the elliptic curve equation. Note that this does not mean that they are the only elements of the set. As a matter of fact, we also include a special point $\mathcal{O}$ and refer to it as the point at infinity. We will motivate the introduction of $\mathcal{O}$ in the next section. For $A, B \in \mathbb{R}$ such that $4A^3 + 27B^2 \neq 0$, the underlying set of the group takes the form:

$$E = \{(x, y) \in \mathbb{R}^2 \mid (y^2 = x^3 + Ax + B)\} \cup \{\mathcal{O}\}.$$

We still need to define a suitable binary operation that acts on a not-necessarily distinct pair of points in E. Any group must satisfy the closure axiom and so the output of the binary operation must also be a point in E. Intuitively, the most natural way of geometrically linking two points in the euclidean plane is with a straight line. It is hence reasonable to look at the different configurations of pairs of points on an elliptic curve defined on $\mathbb{R}$. The diagram below summarizes the possible scenarios:



It is easy to observe that any two elliptic curve points must belong to one of these categories. As a result, these categories are commonly exhaustive. The configurations are also mutually exclusive. This should be clear except possibly for the case of an inflection point. In what follows we argue that an elliptic curve point with ordinate equals to 0 cannot be an inflection point.

0-ordinate points vs. inflection points An inflection point (x_i, y_i) of a curve $y = f(x)$ is one where the *curvature* changes sign. Without delving deeper into the notion of curvature, this means that the second derivative of y with respect to x (assuming it exists on a neighborhood of x_i) changes sign as the x values cross x_i. Intuitively, this suggests the following necessary conditions for a point on the curve (where the second derivative is defined) to be an inflection point:

- The second derivative $\frac{d^2(y)}{dx^2}(x_i, y_i)$ evaluated at (x_i, y_i) is equal to 0, and
- $sign\ [\frac{d^2(y)}{dx^2}(x_i - \epsilon, f(x_i - \epsilon)] \neq sign\ [\frac{d^2(y)}{dx^2}(x_i + \epsilon, f(x_i + \epsilon)]$, for ϵ infinitesimally small.

However, an inflection point could still exist even when the second derivative is not defined at that point. The definition remains the same, i.e., an inflection point is one that marks a change in the curve's concavity. As an example, one can look at the function $y = \sqrt[3]{x}$ defined on $\mathbb{R}$, and verify that the point $(0, 0)$ is an inflection point despite the fact that $\frac{d^2y}{dx^2}$ is not defined at $x = 0$.

The domain of definition of an elliptic curve over $\mathbb{R}$ consists of the set

$$\mathcal{D} \equiv \{x \in \mathbb{R} \mid x^3 + Ax + B \geq 0\}$$

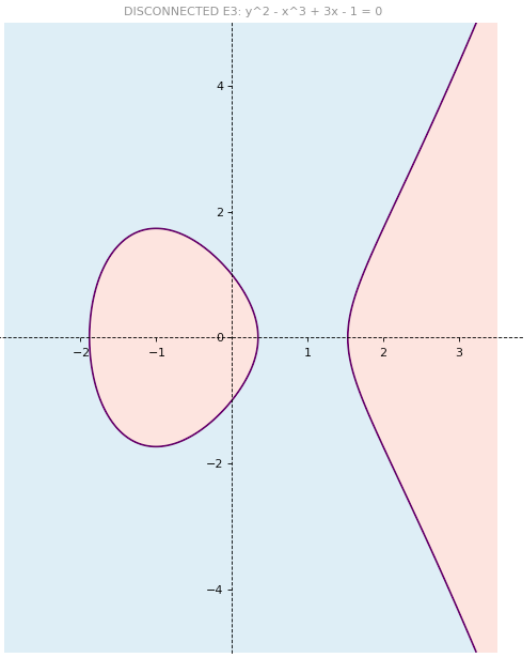
This is due to the fact that over the field of real numbers, square values must be non-negative. Consequently, $y^2 = x^3 + Ax + B$ must be greater than or equal to 0. It then holds that $y = \pm\sqrt{x^3 + Ax + B}$ on $\mathcal{D}$. As a result:

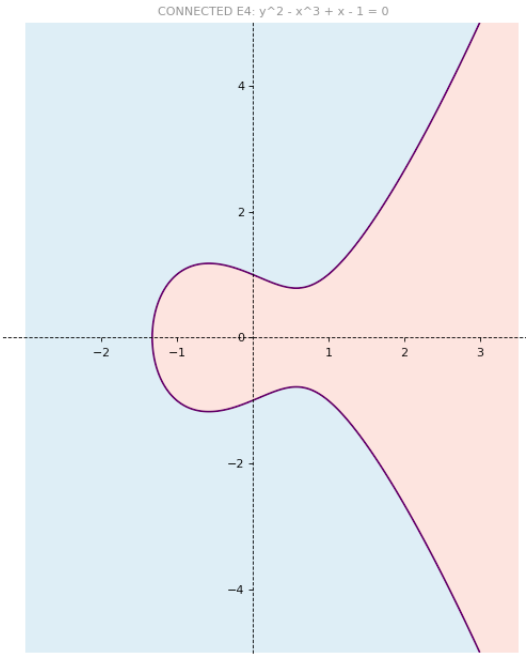
$$\frac{d^2y}{dx^2} = \pm[\frac{-(3x^2+A)^2}{4\sqrt{(x^3+Ax+B)^3}} + \frac{3x}{\sqrt{x^3+Ax+B}}]$$

The second derivative is defined on the set

$$\mathcal{D}^* \equiv \{x \in \mathbb{R} \mid x^3 + Ax + B > 0\}$$

Among other things, this means that the second derivative is **not defined** on curve points whose ordinate is equal to 0. This however, is not enough to justify that 0-ordinate points are not inflection points. To rule out this possibility, we note that by virtue of being a cubic equation, $y = x^3 + Ax + B = 0$ can admit either one or three roots (not necessarily distinct) in $\mathbb{R}$. We can then classify non-singular elliptic curves on $\mathbb{R}$ in two broad categories: **disconnected** or **connected**. The figures below showcase an example of each:





The curve $f(x, y) = y^2 - x^3 + 3x - 1 = 0$ is an example of a non-singular disconnected curve. These curves admit three distinct real roots, none of which are interior points of the domain of definition $\mathcal{D}$ (we do not prove this statement). They are boundary points and hence cannot be crossed from right to left or vice-versa. The same applies to the connected curve $f(x, y) = y^2 - x^3 + x - 1 = 0$ which admits one real root instead of three, but whose unique root is also a boundary point of $\mathcal{D}$.

An implication of the aforestated definition is that the abscissa of an inflection point of a curve in $\mathbb{R}^2$ must be an interior point of $\mathcal{D}$. Indeed, one should be able to cross it in order to validate a change in curvature. Consequently, points on the elliptic curve of the form $(x, 0)$ cannot be inflection points since their abscissas (i.e., the real roots of $x^3 + Ax + B = 0$), are boundary points of $\mathcal{D}$.

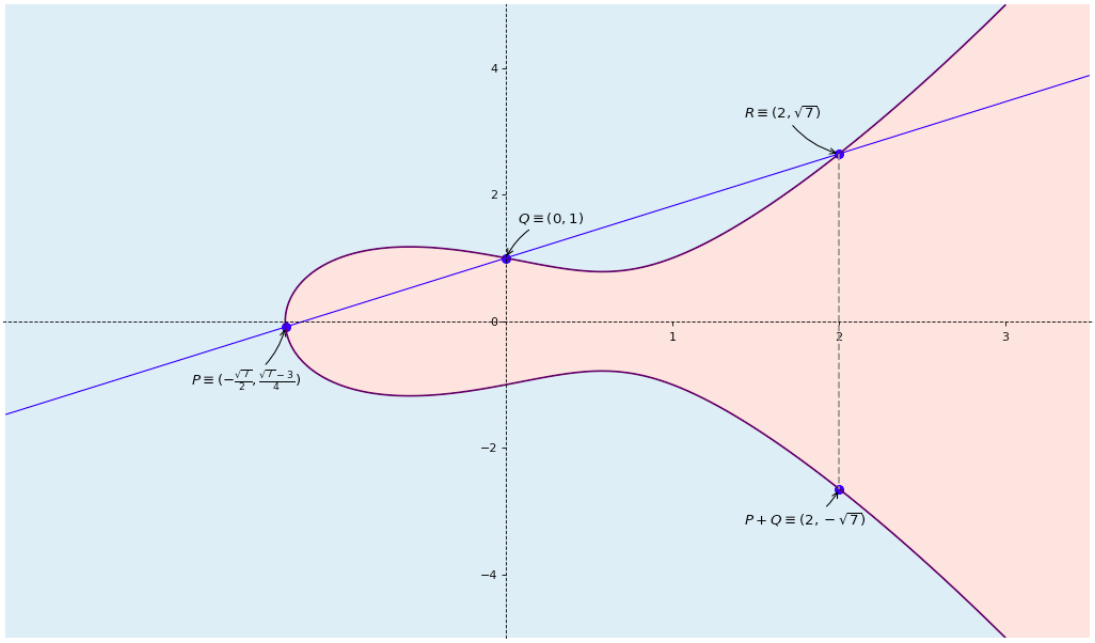
Building the binary operation Having defined the possible configurations of a pair of points on a non-singular elliptic curve on $\mathbb{R}$, we now focus on finding a suitable binary operation. More specifically, given two points on the curve (not-necessarily distinct), our objective is to operate on them in such a way that the output is also a point on the curve.

A rather natural way of doing so is to check if the line passing through the two points intersects the curve at another point. In what follows, we consider each configuration separately and demonstrate that the procedure outputs one or two suitable candidate points. We then show that only one of the two points is permissible (whenever they co-exist), paving the way to an algebraic description of the binary operation.

Configuration #1: $P \equiv (x_p, y_p)$ and $Q \equiv (x_q, y_q)$ are two distinct points on the elliptic curve such that $x_p \neq x_q$.

- The equation of the line joining P and Q is given by $y = (\frac{y_q - y_p}{x_q - x_p})x + (\frac{x_q y_p - x_p y_q}{x_q - x_p})$. We let $c = \frac{y_q - y_p}{x_q - x_p}$, $d = \frac{x_q y_p - x_p y_q}{x_q - x_p}$ and write $y = cx + d$
- As a result, any point of intersection of this line with the elliptic curve must have an abscissa that satisfies $x^3 + Ax + B - (cx + d)^2 = 0$.
- This is a cubic polynomial on $\mathbb{R}$ and as noted earlier, can have either one or three real roots (not necessarily distinct).
- We already know that x_p and x_q are two distinct real roots. Consequently, there must exist a third root (not necessarily distinct from the other two) that we denote by x_{p+q}.
- We write $(x - x_p)(x - x_q)(x - x_{p+q}) = x^3 + Ax + B - (cx + d)^2$, which upon expansion and coefficient comparison shows that $x_{p+q} = c^2 - (x_p + x_q)$.
- Consequently, the point $(x_{p+q}, y_{p+q}) \equiv (c^2 - (x_p + x_q), c^3 - c(x_p + x_q) + d)$ is on the curve.
- Due to $x-$axis symmetry, $(x_{p+q}, -y_{p+q}) \equiv (c^2 - (x_p + x_q), -c^3 + c(x_p + x_q) - d)$ is also on the curve. Later, we specify which of the two points is the eligible one.
- Here is the graph of an instance of this configuration for the following elliptic curve

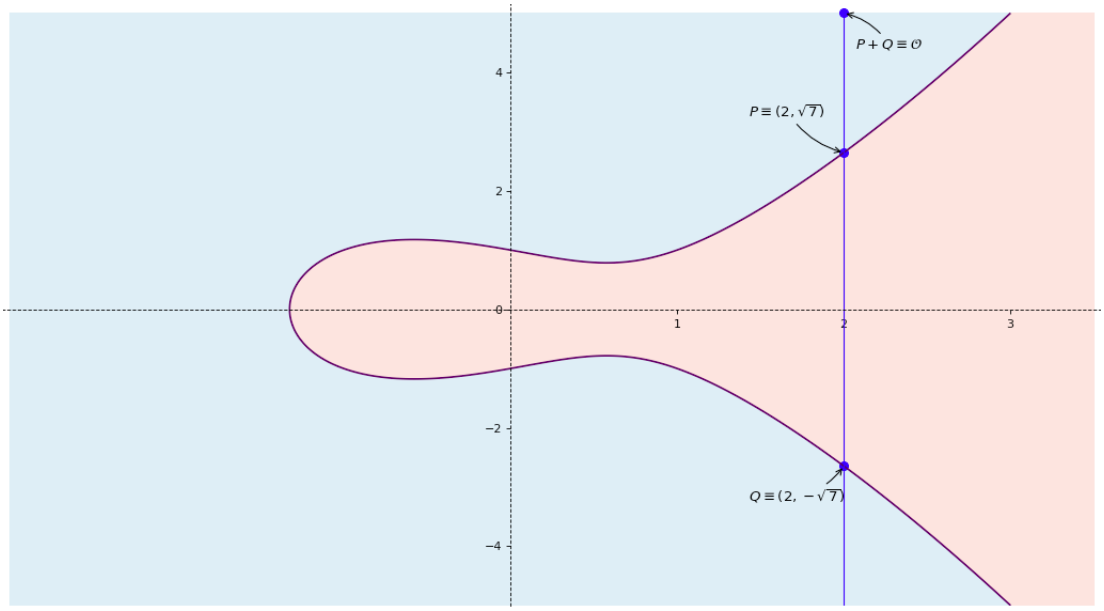
$$E = \{(x, y) \in \mathbb{R}^2 \mid (y^2 - x^3 + x - 1 = 0)\} \cup \{\mathcal{O}\}.$$



Configuration #2: $P \equiv (x_p, y_p)$ and $Q \equiv (x_q, y_q)$ are two distinct points on the elliptic curve such that $x_p = x_q$ (and hence $y_p \neq y_q$).

- The equation of the line joining P and Q is given by $x = x_p = x_q$.
- As a result, any point of intersection of this line with the elliptic curve must have an ordinate that satisfies the equation $y^2 - (x_p^3 + Ax_p + B) = 0$.
- This is a quadratic polynomial in y on $\mathbb{R}$. It either has no solution or admits two not-necessarily distinct real roots.
- We already know that y_p and y_q are two distinct roots, and so we concude that these are the only solutions to the equation.
- The logic of our construction requires that a viable binary operation involving two roots yield a third one. Consequently, we add the point at infinity to the set of eligible group elements and let $P + Q = \mathcal{O}$. In this case we only have one elgible candidate as opposed to two.
- Here is the graph of an instance of this configuration for the following elliptic curve

$$E = \{(x, y) \in \mathbb{R}^2 \mid (y^2 - x^3 + x - 1 = 0)\} \cup \{\mathcal{O}\}.$$



Configuration #3: The two points on the elliptic curve are identical and have an ordinate equal to 0. We let the point be denoted by $P \equiv (x_p, 0)$.

- In the case of two distinct points, there was one and only one line that connected them. A single point however, admits an infinity of lines that go through it. A choice must hence be made.
- To do so, imagine that Q is a distinct point on the curve that is infinitesimally close to P. In the limit as Q approaches P, the line connecting them converges to the line of choice, i.e., the tangent to the elliptic curve at P.

- Note that on the domain $\mathcal{D}$ of the elliptic curve, $y = \pm\sqrt{x^3 + Ax + B}$. Consequently, $\frac{dy}{dx} = \pm\frac{3x^2+A}{2\sqrt{x^3+Ax+B}}$ on $\mathcal{D}^*$. The tangent to the elliptic curve at a 0-ordinate point is a vertical line. In particular, the equation of the tangent at $P \equiv (x_p, 0)$ is given by $x = x_p$.

- Any point of intersection of this tangent with the elliptic curve must have an ordinate that satisfies the equation $y^2 - (x_p^3 + Ax_p + B) = 0$.

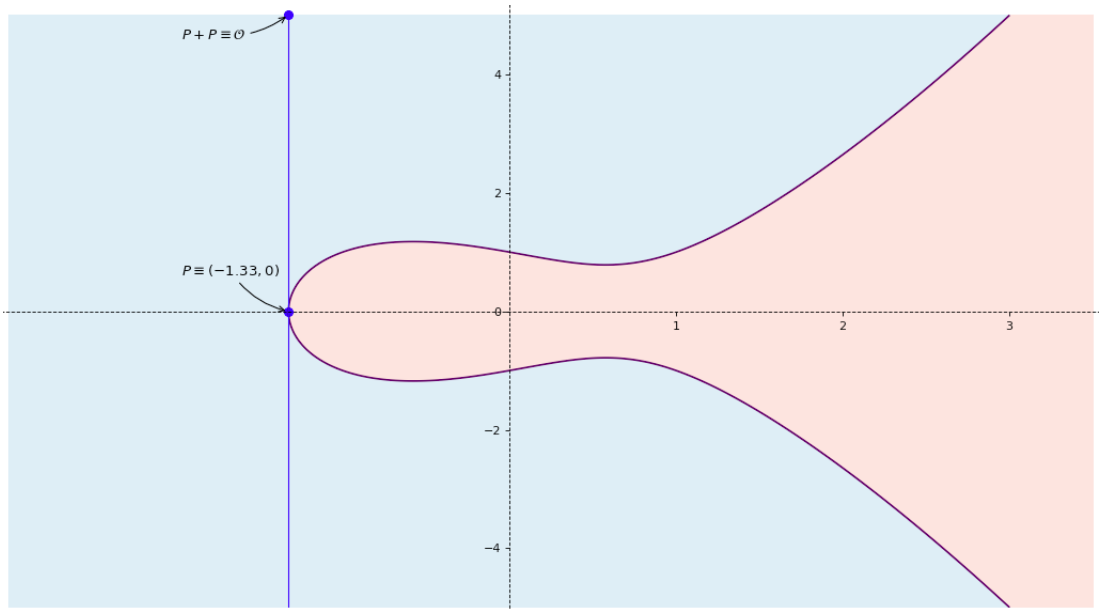
- This is a quadratic polynomial in y on $\mathbb{R}$. It either has no solution or admits two not-necessarily distinct real roots.

- We already know that $y_p = 0$ is a root, hence so is $-y_p = 0$. Consequently 0 is a double root and as a result, the only solution to the equation.

- Following the same logic of configuration #2, we make use of $\mathcal{O}$ (the point at infinity) and let $P + Q = \mathcal{O}$. In this case too, we only have one eligible candidate as opposed to two.

- Here is the graph of an instance of this configuration for the following elliptic curve

$$E = \{(x, y) \in \mathbb{R}^2 \mid (y^2 - x^3 + x - 1 = 0)\} \cup \{\mathcal{O}\}.$$



Configuration #4: The two points on the elliptic curve are identical and constitute an inflection point. We let the point be denoted by $P \equiv (x_p, y_p)$.

- Following the same logic of configuration #3, we choose the line tangent to the elliptic curve at P.

- Without loss of generality let $y = \sqrt{x^3 + Ax + B}$ on $\mathcal{D}$. Consequently, $\frac{dy}{dx} = \frac{3x^2+A}{2\sqrt{x^3+Ax+B}}$ on $\mathcal{D}^*$.

- Let $z \equiv 3x^2 + A$. Hence $\frac{dy}{dx} = \frac{z}{2y}$. The equation of the tangent at P becomes $y = \frac{z_p}{2y_p}x + y_p - \frac{z_p}{2y_p}x_p$

- Any point of intersection of this tangent with the elliptic curve must have coordinates (x, y) that satisfy $4y_p^2y^2 = (z_px + 2y_p^2 - z_px_p)^2$. This can be simplified to $(x - x_p)(z_p^2(x - x_p) + 4y_p^2z_p) = 4y_p^2(y^2 - y_p^2)$.

- We claim that this equation admits x_p as a triple root. To see why, note that $\frac{d^2y}{dx^2} = \frac{-(3x^2+A)^2}{4\sqrt{(x^3+Ax+B)^3}} + \frac{3x}{\sqrt{x^3+Ax+B}}$ over $\mathcal{D}^*$. This can be written as $\frac{d^2y}{dx^2} = -\frac{z^2}{4y^3} + \frac{3x}{y}$.

- A necessary condition for a point on the curve (where the second derivative is defined) to be an inflection point is that its second derivative be 0. Since P is such a point, it must be that $12x_py_p^3 = y_pz_p^2$. And since $y_p \neq 0$, we get $12x_py_p^2 = z_p^2$.

- Substituting z_p^2 with $12x_py_p^2$ in $(x - x_p)(z_p^2(x - x_p) + 4y_p^2z_p) = 4y_p^2(y^2 - y_p^2)$ yields $(x - x_p)(12x_py_p^2(x - x_p) + 4y_p^2z_p) = 4y_p^2(y^2 - y_p^2)$. Since $y_p \neq 0$, we can cancel the $4y_p^2$ factor from both sides and obtain $(x - x_p)(3x_px - 3x_p^2 + z_p) = (y^2 - y_p^2)$.

- Substituting z_p with $3x_p^2 + A$, and y^2, y_p^2 with their x and x_p expressions, we get $(x - x_p)(3x_px - 3x_p^2 + 3x_p^2 + A) = x^3 + Ax + B - x_p^3 - Ax_p - B$. This is equivalent to $(x - x_p)(3x_px) + A(x - x_p) = x^3 - x_p^3 + A(x - x_p)$.

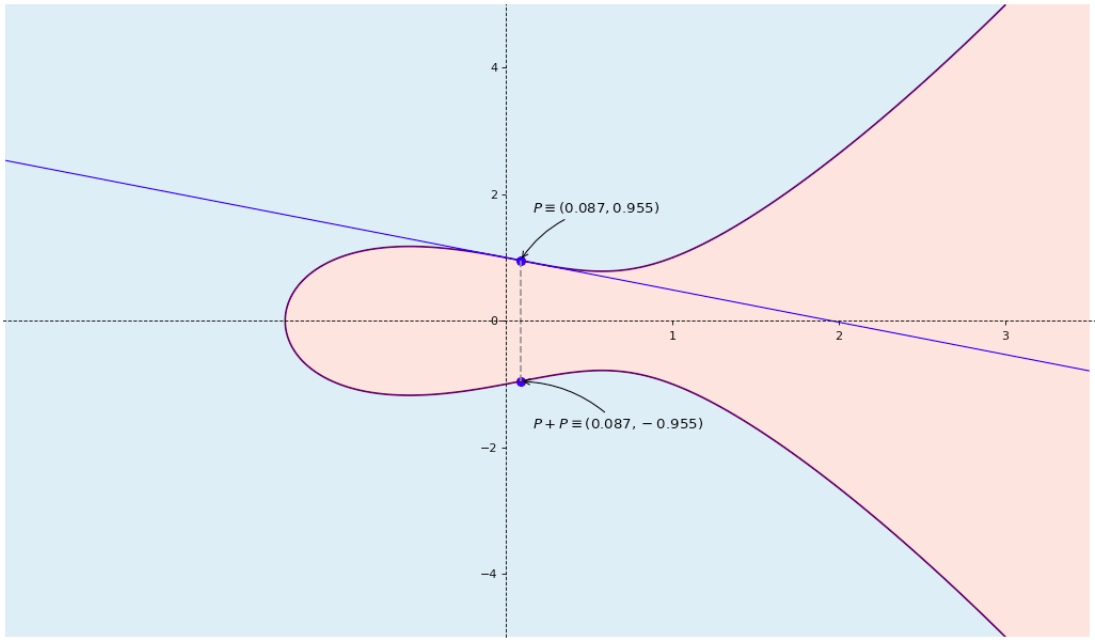
- Canceling $A(x - x_p)$, we get $x^3 - 3x^2x_p + 3xx_p^2 - x_p^3 = 0$. This is equivalent to $(x - x_p)^3 = 0$. Consequently, x_p is a triple root.

- Since a cubic equation has a maximum of three roots on $\mathbb{R}$, x_p is the only one.

- As a result, we could define a binary operation in one of two natural ways: $P + P = P$ or $P + P = \overline{P} \equiv (x_p, -y_p)$. Shortly, we will see why we choose the latter.

- Here is the graph of an instance of this configuration for the following elliptic curve

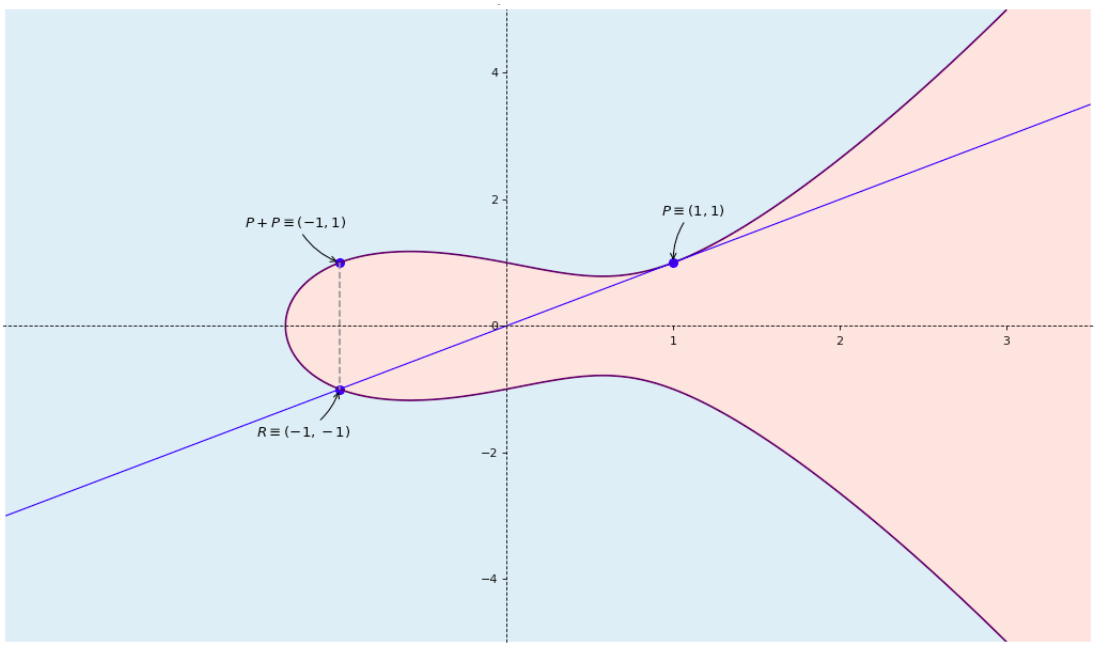
$$E = \{(x, y) \in \mathbb{R}^2 \mid (y^2 - x^3 + x - 1 = 0)\} \cup \{\mathcal{O}\}.$$



Configuration #5: The two points on the elliptic curve are identical, have non-zero ordinate and are not an inflection point. We let the point be denoted by $P \equiv (x_p, y_p)$.

- As in configuration #3, we choose the line tangent to the elliptic curve at P.
- On $\mathcal{D}^*$, the tangent's equation at P is $y = \frac{3x_p^2+A}{2y_p}x + y_p - \frac{3x_p^2+A}{2y_p}x_p$. Letting $c = \frac{3x_p^2+A}{2y_p}$ and $d = y_p - \frac{3x_p^2+A}{2y_p}x_p$, we get $y = cx + d$.
- As a result, any point of intersection of this line with the elliptic curve must have an abscissa that satisfies $x^3 + Ax + B - (cx + d)^2 = 0$
- This is a cubic polynomial on $\mathbb{R}$ and as noted earlier, can have either one or three real roots (not necessarily distinct)
- We know that x_p is a double real root, and so there must be a third root. We denote it x_{p+p}.
- We write $(x - x_p)(x - x_p)(x - x_{p+p}) = x^3 + Ax + B - (cx + d)^2$, which upon expansion and coefficient comparison yields $x_{p+p} = c^2 - 2x_p$.
- As a result, $(x_{p+p}, y_{p+p}) \equiv (c^2 - 2x_p, c^3 - 2cx_p + d)$ is on the curve. Due to $x-$axis symmetry, $(x_{p+p}, -y_{p+p}) \equiv (c^2 - 2x_p, -c^3 + 2cx_p - d)$ is also on the curve.
- Note that in configuration #1, if the resulting third root were equal to either x_p or x_q (recall that P and Q have distinct absiscas), the outcome would be similar to that of configuration #5.
- Here is the graph of an instance of this configuration for the following elliptic curve

$$E = \{(x, y) \in \mathbb{R}^2 \mid (y^2 - x^3 + x - 1 = 0)\} \cup \{\mathcal{O}\}.$$



Choosing a candidate In configurations #1, #4 and #5, we ended-up with two points (symmetric about the x-axis) to choose from with regard to the output of the binary operation. Only one of them safeguards the group axioms. To see which one does not, consider configuration #4:

- Suppose that we choose point P (as opposed to $\overline{P}$). That means that $P + P = P$.
- Consequently, P must be the identity of the group. We claim that this can not be.
- To see why, choose a curve point $Q \neq P$ such that (P, Q) is not tangent to the curve at Q.
- One can see that this is a special case of configuration #1 (if $x_p \neq x_q$) or of configuration #2 (if $x_p = x_q$), and that the procedure previously described results in candidates all of which are different than Q.
- As a result, P cannot be the identity element and is thus ruled out.

Defining the elliptic curve group We are now in a position to introduce the elliptic curve group $(E(\mathbb{R}), \oplus)$ on $\mathbb{R}$ and verify that it respects the abelian group axioms.

- For given $A,\ B \in \mathbb{R}$ such that $4A^3 + 27B^2 \neq 0$, the underlying set of the group is defined to be:

$$E(\mathbb{R}) = \{(x, y) \in \mathbb{R}^2 \mid (y^2 = x^3 + Ax + B)\} \ \cup \ \{\mathcal{O}\}.$$

- Let $P \equiv (x_p, y_p)$ and $Q \equiv (x_q, y_q)$ be points in $E(\mathbb{R})$. We define the binary operation $\oplus$ as follows:

  { If $P \neq Q$ and $x_p \neq x_q$ (i.e., configuration #1), let $c = \frac{y_q - y_p}{x_q - x_p}$, $d = \frac{x_q y_p - x_p y_q}{x_q - x_p}$,

$$P \oplus Q \equiv (c^2 - (x_p + x_q), -c^3 + c(x_p + x_q) - d)$$

  { If $P \neq Q$ and $x_p = x_q$ (i.e., configuration #2),

$$P \oplus Q \equiv \mathcal{O}$$

  { If $P = Q$ and $y_p = y_q = 0$ (i.e., configuration #3),

$$P \oplus Q = P \oplus P \equiv \mathcal{O}$$

  { If $P = Q$ and $y_p = y_q \neq 0$ (i.e., configurations #4 and #5), let $c = \frac{3x_p^2 + A}{2y_p}$, $d = y_p - \frac{3x_p^2 + A}{2y_p} x_p$,

$$P \oplus Q \equiv (c^2 - 2x_p, -c^3 + 2cx_p - d)$$

- The point at infinity $\mathcal{O}$ is defined to be the identity element of the elliptic group.

$(E(\mathbb{R}), \oplus)$ thus defined, satisfies the abelian group axioms:

1. *Associativity*: Using the aforestated definition of $\oplus$, one can verify (with a bit of patience and willingness to write down lengthy formulas) that

$$P, Q, R \in E(\mathbb{R}) \Rightarrow P \oplus (Q \oplus R) = (P \oplus Q) \oplus R$$

2. *Existence of identity*: We defined $\mathcal{O} \in E(\mathbb{R})$ such that for all $P \in E(\mathbb{R})$, we have

$$P \oplus \mathcal{O} = \mathcal{O} \oplus P = P$$

Note that for any $P, Q \in E(\mathbb{R})$ such that $P \neq \mathcal{O},\ P + Q$ is never equal to Q. This can be readily verified by checking each configuration separately.

3. *Closure*: By construction of $\oplus$, we made sure that the result of adding two points on the elliptic curve is another point on the curve. In other terms

$$P, Q \in\ E(\mathbb{R}) \Rightarrow P \oplus Q \in E(\mathbb{R})$$

4. *Existence of inverse*: $\forall P \equiv (x_p, y_p) \in E(\mathbb{R})$, the point $\overline{P} \equiv (x_p, -y_p)$ is also in $E(\mathbb{R})$. This is because the elliptic curve exhibits x-axis symmetry. Moreover, P and $\overline{P}$ belong to either configuration #2 (if they are distinct) or configuration #3 (if they are identical) and so the definition of $\oplus$ ensures that

$$P \oplus \overline{P} = \overline{P} \oplus P = \mathcal{O}$$

As a result, $\overline{P}$ is the inverse of P.

5. *Commutativity*: The definition of $\oplus$ implies that

$$P, Q \in E(\mathbb{R}) \Rightarrow P \oplus Q = Q \oplus P$$

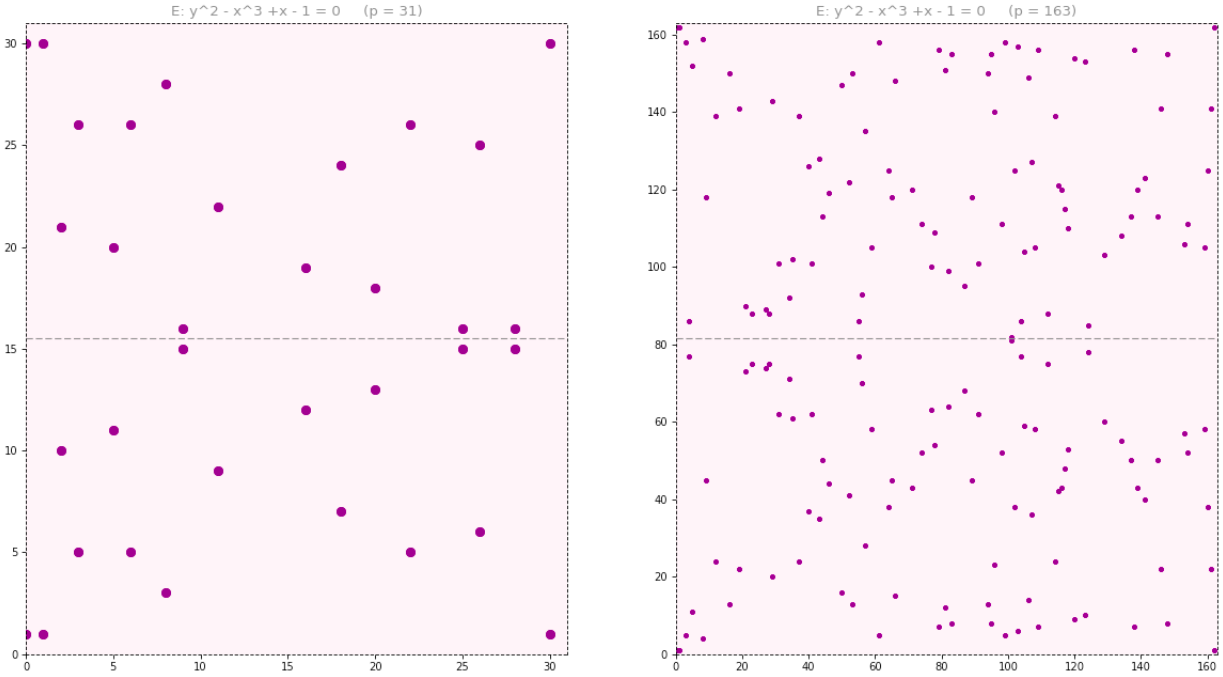
4 Elliptic curve groups over finite fields

Going forward, we only consider finite fields of prime order p and do not cover extension fields. A non-singular elliptic curve E defined over a finite field $\mathbb{F}_p$ of prime order $p \notin \{2, 3\}$ differs from one defined over $\mathbb{R}$ in the following way:

1. The equation of the curve becomes $E : y^2 \equiv x^3 + Ax + B \pmod{p}$ as opposed to $E : y^2 = x^3 + Ax + B$
2. The parameters A and B are chosen in $\mathbb{F}_p$ as opposed to $\mathbb{R}$
3. The discriminant must satisfy $4A^3 + 27B^2 \neq 0 \pmod{p}$ as opposed to $4A^3 + 27B^2 \neq 0$
4. Points on E consist of tuples $(x, y) \in \mathbb{F}_p^2$ such that $y^2 \equiv x^3 + Ax + B \pmod{p}$. This is in contrast to tuples $(x, y) \in \mathbb{R}^2$ such that $y^2 = x^3 + Ax + B$.

The main difference is that all computations are conducted in modulo p arithmetic. In what follows, we depict the elliptic curve $E : y^2 \equiv x^3 - x + 1 \pmod{p}$ over $\mathbb{F}_{31}$ and over $\mathbb{F}_{163}$. The geometry of elliptic curves over finite fields is not as intuitive as that of those

over $\mathbb{R}$. However, we will see that the algebraic formulation of their associated group closely follows that of elliptic groups over $\mathbb{R}$.



For example, in order to draw $E : y^2 \equiv x^3 - x + 1 \pmod{p}$ over $\mathbb{F}_{31}$, we select each value of x in the set $\{0, 1, 2, ..., 30\}$ and plug it into the expression $x^3 - x + 1 \pmod{31}$. We subsequently check whether the result is a quadratic residue or not by verifying whether $\exists y \in \{0, 1, 2, ..., 30\}$ such that $y^2 \pmod{31}$ is a match. We find that the euclidean representation of this curve over $\mathbb{F}_{31}$ consists of the following 34 elements:

$$\{(0,1),(0,30),(1,1),(1,30),(2,10),(2,21),(3,5),(3,26),(5,11),(5,20),(6,5),(6,26),(8,3),(8,28),(9,15),(9,16),(11,9),(11,22),(16,12),(16,19),(18,7),(18,24),(20,13),(20,18),(22,5),(22,26),(25,15),(25,16),(26,6),(26,25),(28,15),(28,16),(30,1),(30,30)\}.$$

While on $\mathbb{R}$ the elliptic curve exhibited x-axis symmetry, on $\mathbb{F}_p$ it exhibits symmetry about the horizontal line $y = \frac{p}{2}$. Indeed, if (x, y) is a point on the curve, then so will $(x, -y + p)$. This is because $(-y + p)^2 \pmod{p} \equiv y^2 + p^2 - 2yp \pmod{p} \equiv y^2 \pmod{p}$.

Formally, we denote by $(E(\mathbb{F}_p), \oplus_p)$ the group associated with an elliptic curve defined over $\mathbb{F}_p$. In particular:

- For $p \notin \{2, 3\}$ and for given $A,\ B \in \mathbb{F}_p$ such that $4A^3 + 27B^2 \neq 0 \pmod{p}$ the underlying set of the group is defined to be:

$$E(\mathbb{F}_p) = \{(x, y) \in \mathbb{F}_p^2 \mid y^2 \equiv x^3 + Ax + B \pmod{p}\} \cup \{\mathcal{O}\}.$$

- Let $R \equiv (x_r, y_r)$ and $Q \equiv (x_q, y_q)$ be points in $E(\mathbb{F}_p)$. We define the binary operation $\oplus_p$ as follows:

  { If $R \neq Q$ and $x_r \neq x_q \pmod{p}$, let $c \equiv \frac{y_q - y_r}{x_q - x_r} \pmod{p}$, $d \equiv \frac{x_q y_r - x_r y_q}{x_q - x_r} \pmod{p}$ (division refers to multiplication by the inverse of the denominator over $\mathbb{F}_p$),

$$R \oplus_p Q \equiv (c^2 - (x_r + x_q) \pmod p, -c^3 + c(x_r + x_q) - d \pmod p)$$

{ If $R \neq Q$ and $x_r \equiv x_q \pmod p$,

$$R \oplus_p Q \equiv \mathcal{O}$$

{ If $R = Q$ and $y_r \equiv y_q \equiv 0 \pmod p$,

$$R \oplus_p Q = R \oplus_p R \equiv \mathcal{O}$$

{ If $R = Q$, $y_r \neq 0 \pmod p$, let $c \equiv \frac{3x_r^2+A}{2y_r} \pmod p$, $d \equiv y_r - \frac{3x_r^2+A}{2y_r}x_r \pmod p$ (division means multiplication by the inverse of the denominator over $\mathbb{F}_p$),

$$R \oplus_p Q \equiv (c^2 - 2x_r \pmod p, -c^3 + 2cx_r - d \pmod p)$$

- The point at infinity $\mathcal{O}$ is defined to be the identity element of the elliptic group.

To illustrate point addition in elliptic curve groups over finite fields, we look at $(E(\mathbb{F}_{31}), \oplus_{31})$ and operate on points $R \equiv (3, 26)$ and $Q \equiv (28, 15)$. Since $R \neq Q$ and $x_r \neq x_q \pmod{31}$ we compute

- $c \equiv \frac{y_q - y_r}{x_q - x_r} \pmod p \equiv \frac{15-26}{28-3} \pmod{31} \equiv \frac{-11}{25} \pmod{31} \equiv -11 \times 5 \pmod{31} = 7$

  where 5 is the inverse of 25 in modulo 31 arithmetic (recall that this can be efficiently computed using the extended euclidean algorithm introduced in the *Groups and Finite Fields* post).

- $d \equiv \frac{x_q y_r - x_r y_q}{x_q - x_r} \pmod p \equiv \frac{28\times 26 - 3\times 15}{28-3} \pmod{31} \equiv 683 \times 5 \pmod{31} = 5$

We then compute $R \oplus_{31} Q \equiv (7^2 - (3 + 28) \pmod{31}, -7^3 + 7(3 + 28) - 5 \pmod{31}) = (18, 24)$.

The construction of $(E(\mathbb{F}_p), \oplus_p)$ is similar to that of $(E(\mathbb{R}), \oplus)$ except for the fact that values are computed modulo p. $(E(\mathbb{F}_p), \oplus_p)$ thus defined, satisfies the abelian group axioms:

1. *Associativity*: Using the aforestated definition of $\oplus_p$, one can verify after lengthy and tedious calculations that

$$Q, R, S \in E(\mathbb{F}_p) \Rightarrow Q \oplus_p (R \oplus_p S) = (Q \oplus_p R) \oplus_p S$$

2. *Existence of identity*: We defined $\mathcal{O} \in E(\mathbb{F}_p)$ such that for all $Q \in E(\mathbb{F}_p)$, we have

$$Q \oplus_p \mathcal{O} = \mathcal{O} \oplus_p Q = Q$$

3. *Closure*: By definition of $\oplus_p$, the resulting output is either $\mathcal{O}$ or a tuple $(x, y) \in \mathbb{F}_p^2$ (since all arithmetic is conducted modulo p). Moreover, one can readily use the definition of $\oplus_p$ to check that the result of the binary operation always verifies the elliptic curve equation modulo p.. Consequently,

$$Q, R \in E(\mathbb{F}_p) \Rightarrow Q \oplus_p R \in E(\mathbb{F}_p)$$

4. *Existence of inverse*: $\forall Q \equiv (x_q, y_q) \in E(\mathbb{F}_p)$, the point $\overline{Q} \equiv (x_q, -y_q + p)$ is also in $E(\mathbb{F}_p)$ due to symmetry about the line $y = \frac{p}{2}$. And so the definition of $\oplus_p$ ensures that

$$Q \oplus_p \overline{Q} = \overline{Q} \oplus_p Q = \mathcal{O}$$

   As a result, $\overline{Q}$ is the inverse of Q.

5. *Commutativity*: The definition of $\oplus_p$ implies that

$$Q, R \in E(\mathbb{F}_p) \Rightarrow Q \oplus_p R = R \oplus_p Q$$

ECDLP, cardinality, and point multiples in $(E(\mathbb{F}_p), \oplus_p)$. Recall that the importance of elliptic curve groups over finite fields is largely derived from the exponential hardness of the DL problem on them. The Elliptic Curve Discrete Logarithm Problem (also known as **ECDLP**) can be stated as follows:

- Let E be an elliptic curve defined over a finite field $\mathbb{F}_p$ (i.e., $A, B \in \mathbb{F}_p$ such that $4A^3 + 27B^2 \neq 0 \pmod p$)

$$E : y^2 \equiv x^3 + Ax + B \pmod p$$

- Let $(E(\mathbb{F}_p), \oplus_p)$ be the group associated with it, where

$$E(\mathbb{F}_p) \equiv \{(x, y) \in \mathbb{F}_p^2 \mid y^2 \equiv x^3 + Ax + B \pmod p\} \cup \{\mathcal{O}\}.$$

- Let $Q, R \in E(\mathbb{F}_p)$ and find the smallest integer m (if it exists) such that

$$R = m \otimes_p Q \equiv Q \oplus_p Q \oplus_p \ ... \ \oplus_p Q \ (m \text{ times})$$

The notation $m \otimes_p Q$ is unusual as it is commonly written as mQ. We decide to make explicit the appearance of the operator $\otimes_p$ as a reminder that it is scalar multiplication with respect to the binary operator $\oplus_p$.

Finding such an m when it exists is thought to be exponentially hard. In the context of crypto-assets, we don't operate on the full set $E(\mathbb{F}_p)$. Rather, we choose an element $G \in E(\mathbb{F}_p)$ such that order(G) is a very large prime. We then limit ourselves to the subgroup $(\{G\}, \oplus_p)$ generated by G (refer to the post on *Groups and Finite Fields* for an introduction to subgroups). Given G and $M \in \{G\}$, ECDLP now consists in finding the smallest integer m such that $M = m \otimes_p G$. We are confident that such an m exists since $M \in \{G\}$.

In the digital signature schemes used in e.g., Bitcoin and Monero, m represents the private key and M the public one. It is important to derive M efficiently from m. However, as we stated earlier, it must not be polynomially feasible to compute m from M. The exponential hardness of ECDLP helps with the latter requirement. Moreover, one can expect that the larger the set $E(\mathbb{F}_p)$, the better. This justifies the importance of having a sense of the cardinality of $E(\mathbb{F}_p)$, also denoted $\#E(\mathbb{F}_p)$. In order to ensure

the former requirement, we need to have an efficient polynomial-time algorithm that can compute multiples of G.

1. **Cardinality of $E(\mathbb{F}_p)$** We previously saw (e.g., in the case of $E(\mathbb{F}_{31})$) that not every tuple $(x, y) \in \mathbb{F}_p^2$ is necessarily an element of $E(\mathbb{F}_p)$. To get an upper-bound on $\#E(\mathbb{F}_p)$, note that for every $x \in \mathbb{F}_p$ one can have at most two values of y that satisfy the elliptic curve equation $E : y^2 \equiv x^3 + Ax + B \pmod{p}$.

   If y is a solution, then so will $-y + p$. In addition, we have the point at infinity $\mathcal{O}$. As a result, since there are p distinct values in $\mathbb{F}_p$, we get a maximum of $2p + 1$ points in $E(\mathbb{F}_p)$. Over $\mathbb{F}_p,\ p \neq 2$, there are $\frac{p-1}{2}$ quadratic residues and an equal number of quadratic non-residues (we don't prove this statement in this post). As a result, in the absence of any information, a random $x \in \mathbb{F}_p$ has equal probability of being a square or not. One can then calculate the expected value of the number of points in $E(\mathbb{F}_p)$ to be $\frac{2p}{2} + 1 = p + 1$. The German mathematician Helmut Hasse showed that

   $$|\#E(\mathbb{F}_p) - (p + 1)| \leq 2\sqrt{p}$$

   The Dutch mathematician Ren Schoof, relied partly on Hasse's theorem to devise a deterministic algorithm that can compute $\#E(\mathbb{F}_p)$ with complexity $\mathcal{O}(log^9 p)$. This is known as the Schoof algorithm and its proof is beyond the scope of this post (readers interested in learning more about it can consult [9]). The important take-away is that there exists a polynomial-time algorithm to calculate the order of an elliptic curve group over a finite field.

   In so far as the structure of this group is concerned, we mention without proof that $(E(\mathbb{F}_p), \oplus_p)$ is always either cyclic or the product of two cyclic groups.

2. **Point multiplicity** Suppose m is a very large integer, and $G \in E(\mathbb{F}_p)$ is as defined earlier. In order to calculate $m \otimes_p G \equiv G \oplus_p \ ... \ \oplus_p G$ (m times), one can do the following:

   (a) Write m in its base 2 expansion

   $$m = m_0 + m_1.2 + \ ... \ + m_k.2^k,\ m_i \in \{0, 1\} \text{ for } i \in \{0, ..., k-1\},\ m_k = 1$$

   This can be achieved in $\mathcal{O}(log_2(m))$

   (b) Subsequently, write $m \otimes_p G$ as

   $$m \otimes_p G = m_0 \otimes_p G + 2m_1 \otimes_p G + \ ... \ + 2^k m_k \otimes_p G,\ m_i \in \{0, 1\} \text{ for } i \in \{0, ..., k-1\},\ m_k = 1$$

   (c) Since the value $2^k m_k$ is equal to $2^k,\ 2^k m_k \otimes_p G$ can be evaluated by doubling the point G (i.e., $G \oplus_p G$, then $(G \oplus_p G) \ \oplus_p \ (G \oplus_p G)$, ...) a total of k times. In the process, we store the value of 2^i for all $i \in \{0, ..., k\}$ for which $m_i = 1$.

   (d) Putting it altogether, in order to caculate $m \otimes_p G$, one first needs $\mathcal{O}(log_2(m))$ steps to find $m_i,\ i \in \{0, ..., k-1\}$, then a total of k point doublings, and

finally, a worst-case total of k additions. Hence, the worst-case complexity is $\mathcal{O}(log_2(m) + k + k)$. Noting that k is on the order of $log_2(m)$, the complexity becomes $\mathcal{O}(3 \times log_2(m)) = \mathcal{O}(log_2(m))$.

The above procedure is known as the **double-and-add** method and ensures that point addition can be efficiently carried on elliptic curves over finite fields.

3. **A note on finding** G A question of practical importance is whether one can efficiently find an adequate point $G \in E(\mathbb{F}_p)$ such that its order is equal to a large prime. This question is justified since finite fields of interst have an astronomically large characteristic (usually larger that 2^{160}) and one cannot expect to conduct an exhaustive search on the underlying group. The answer turns out to be positive, and one way of finding a suitable G is as follows:

   3.1 Calculate $\#E(\mathbb{F}_p)$ using Schoof's polynomial-time algorithm.

   3.2 Find a very large prime (ideally the largest) n that divides $\#E(\mathbb{F}_p)$. Say $\#E(\mathbb{F}_p) = n \times h$. Note that the integer factorization problem is thought to be hard. However, in some cases where the integer exhibits certain properties, one can employ more efficient algorithms to identify a prime factor. We will not go into the details of these algorithms but refer the interested reader to e.g., [4] for an accessible overview.

   3.3 Choose a random point $Q \in E(\mathbb{F}_p)$. We know that order(Q) = order$(\{Q\})$. By Lagrange's theorem, the order of any subgroup divides the order of the parent group. As a result, the order of Q must divide $\#E(\mathbb{F}_p)$, and so

$$\mathcal{O} = \#E(\mathbb{F}_p) \otimes_p Q = (n \times h) \otimes_p Q = n \otimes_p (h \otimes_p Q).$$

   3.4 Consequently, the order of the element $h \otimes_p Q$ must divide n. Since n is prime, the order of $h \otimes_p Q$ can either be equal to 1 or n. If it is 1, then $h \otimes_p Q$ is the identity element $\mathcal{O}$. In this case, we go back to step 3.3 and choose a different point Q. Otherwise, we set $G \equiv h \otimes_p Q$ which can be computed efficiently using the aforementioned double-and-add method. We refer to h as the **cofactor** of G.

5 Elliptic curve in Bitcoin

Bitcoin's cryptography relies on a particular curve known as **secp256k1**:

- **"sec"** is short for **Standard for Efficient Cryptography**. It refers to a set of standards developed and published by the *Certicom Research* consortium [7].
- **"p"** refers to the fact that the curve is defined over a finite field $\mathbb{F}_p$ of prime order p.
- **"256"** means that the curve points' abscissas and ordinates are 256-bit long.
- **"k"** states that the curve belongs to the Koblitz family. Usually, Koblitz curves are defined over extension fields of characteristic 2 i.e., over $\mathbb{F}_{2^k}$ for some positive integer k. However, the Bitcoin curve is defined over a prime field of characteristic

$p \neq 2$. It belongs to a more general version of Koblitz curves. Without going into further details, it suffices to say that for all practical matters, the importance of this class of curves is derived from its higher efficiency in computing point multiples on its associated group.

- **"1"** is a reminder that it is the first curve of its kind that satisfies the previous attributes.

The parameters of secp256k1 can be found on page 9 of [7] and are as follows:

- $p = 2^{256} - 2^{32} - 2^9 - 2^8 - 2^7 - 2^6 - 2^4 - 1$, or in hexadecimal notation (hex)

  FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF
  FFFFFFFE FFFFFC2F

  Each element represents a half-byte (i.e., 4 bits) known as a **nibble**. There are 64 nibbles corresponding to the 256-bit representation mandated by the standard.

- $A = 0$, which in standard hex is given by

  00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000

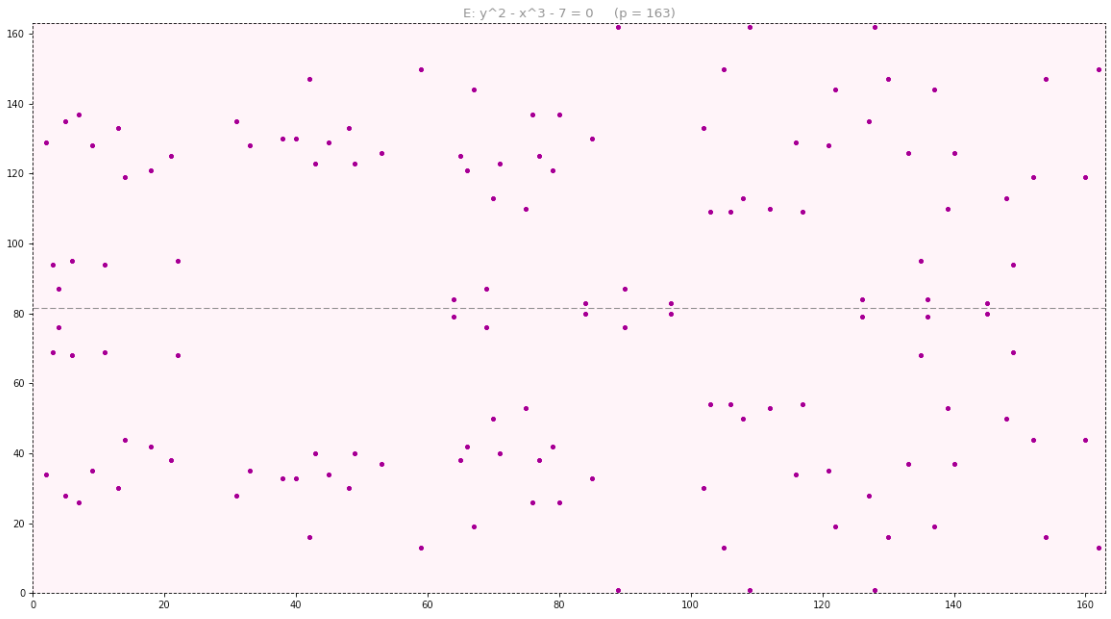
- $B = 7$, which in standard hex is given by

  00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000007

- Since $p \notin \{2, 3\}$, and $4A^3 + 27B^2 \neq 0 \pmod{p}$, the curve is non-singular and can be written in short Weierstrass form. The resulting group is $(E(\mathbb{F}_p), \oplus_p)$, where

$$E(\mathbb{F}_p) = \{(x, y) \in \mathbb{F}_p^2 \mid y^2 \equiv x^3 + 7 \pmod{p}\} \cup \{\mathcal{O}\}.$$

  Here is a euclidean representation of this curve when $p = 163$ *(it is not feasible to show it for* $p = 2^{256} - 2^{32} - 2^9 - 2^8 - 2^7 - 2^6 - 2^4 - 1$*).*



- The base point G has abscissa and ordinate given by

$$x_G \equiv 55066263022277343669578718895168534326250603453777594175500187360389116729240 \pmod{p}$$

$$y_G \equiv 32670510020758816978083085130507043184471273380659243275938904335757337482424 \pmod{p}$$

  which in standard hex notation are given by:

$$x_G = \text{79BE667E F9DCBBAC 55A06295 CE870B07 029BFCDB 2DCE28D9 59F2815B 16F81798}$$

$$y_G = \text{483ADA77 26A3C465 5DA4FBFC 0E1108A8 FD17B448 A6855419 9C47D08F FB10D4B8}$$

  Bitcoin's public-key cryptography is hence conducted on the subgroup $(\{G\}, \oplus_p)$.

- The order of G is chosen to be a prime number equal to

$$n = 115792089237316195423570985008687907852837564279074904382605163141518161494337 \pmod{p}$$

  which in standard hex notation is given by

$$n = \text{FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFE BAAEDCE6 AF48A03B BFD25E8C D0364141}$$

- Recall that n denotes the order of G, and must divide $\#E(\mathbb{F}_p)$ i.e., the order of $E(\mathbb{F}_p)$. The cofactor h is equal to $\frac{\#E(\mathbb{F}_p)}{n}$, which in this case is equal to 1 and is represented in standard hex notation by

  $n = \text{0000000 00000000 00000000 00000000 00000000 00000000 00000000 00000001}$

  That means that the order of G is equal to that of $E(\mathbb{F}_p)$, i.e., $n = \#E(\mathbb{F}_p)$. Since n is prime, the order of $E(\mathbb{F}_p)$ must also be prime. As a result, $(E(\mathbb{F}_p), \oplus_p)$ is a cyclic group and any of its elements could serve as a generator (refer to *Groups and Finite Fields* for an introduction to cyclic groups).

Another noteworthy SEC2 curve is **secp256r1**. The **'r'** specifier refers to the attribute "random" since the generation of the curve parameters A and B relies on a supposedly random process involving a seed value fed to a hash function. The seed value as well as the other curve attributes can be found on pages 9 and 10 of [7].

There was a fair amount of questioning as to why Satoshi opted for the usage of secp256k1 as opposed to that of another curve such as secp256r1. The reason(s) remain obscure and advocates that favor one curve over the other abound (e.g., [5], [1]). The point of contention lies in the randomness involved in selecting the curve parameters:

- On the one hand, Koblitz curves exhibit slightly weaker security than other curves (although when using 256-bit long parameters, the difference is negligible). Moreover, the National Institute of Standards and Technology or NIST (a US governmental agency) has been advocating the usage of secp256r1 on the basis that its parameters achieve very high security standards. Another common name of this curve is **NIST P-256**, and constitutes one of fifteen curves that NIST recommends.
- On the other hand, skeptics argue that NIST's endorsement, coupled with the absence of a rationale for the choice of the seed value are ground for dismissal. According to them, it is possible that the NIST (and affiliated entities) might have placed a backdoor to weaken the curve's security standard. Their scepticism is not unfounded since the NSA and NIST have previously planted a backdoor in the elliptic curve algorithm known as Dual_EC_DRBG, which was validated by memos leaked by Edward Snowden. The interested reader can refer to [11] for a take on how the NSA might have accomplished this.

Suffice it to say that no one can tell with certainty whether one curve is preferred over the other. Assuming no backdoor, both curves exhibit comparable security standards.

6 Elliptic curve in Monero

The NIST debacle surrounding the Dual_EC_DRBG algorithm pushed some people away from NIST curves and closer to curves generated in academic circles instead. Two such curves are **Curve25519** and its next of kin **ed25519** used in Monero. Both are elliptic curves, but are not represented in short Weierstrass form. However, they could be transformed into one and we will see how shortly.

Curve25519 was originally introduced by the German-American mathematician and cryptologist Daniel Julius Bernstein. Unlike SEC curves and some of those advocated by NIST, Curve25519 is thougt to be patent-free. It is also hailed for its faster computation of point multiples when compared to e.g., sec256r1 (NIST P-256) [6]. Moreover, it exhibits a security level comparable to that of secp256k1 and secp256r1 (assuming no backdoors). These favorable attributes paved the way for its ever-increasing adoption.

We first provide an overview of **Montgomery** and **Twisted Edward** representations of elliptic curves of which Curve25519 and ed25519 are respective examples. We show that under certain constraints, any of these representations could be transformed into a short Weierstrass counterpart using a specific **isomorphism**. The existence of an isomorphism makes the two curves' respective groups **equivalent** and guarantees that the hardness of ECDLP is preserved on both. In the last section, we introduce the attributes of Monero's ed25519 curve.

Twisted Edward and Montgomery representations A Twisted Edward curve defined on a field $\mathbb{K}$ such that $char(\mathbb{K}) \neq 2$ with parameters $a \in \mathbb{K}$ and $d \in \mathbb{K}$ such that $ad(a-d) \neq 0$, is one that satisifies the following equation

$$E^E_{a,d} : ax^2 + y^2 = 1 + dx^2y^2$$

It turns out that if in addition, a is a square in $\mathbb{K}$ and d is not, the curve will define a group structure. For our purposes, on a finite field $\mathbb{F}_p,\ p > 2$, such a curve will define a group $(E^E_{a,d}(\mathbb{F}_p), \oplus^E_p)$ where the superscript E refers to "Edward". One could define the binary operation $\oplus^E_p$ from basic principles as we did earlier for curves in short Weierstrass form. However, we show shortly that each such Twisted Edward curve is equivalent to another one in short Weierstrass form. The equivalence implicitly defines a corresponding group structure associated with it. The underlying set of the group is defined as

$$E^E_{a,d}(\mathbb{F}_p) \equiv \{(x,y) \in \mathbb{F}^2_p \mid ax^2 + y^2 \equiv 1 + dx^2y^2 \pmod p\}$$

Note that it does not contain a point at infinity. We will attempt to justify its absence when we discuss the equivalence betwen curve representations.

A Montgomery curve defined on a field $\mathbb{K}$ such that $char(\mathbb{K}) \neq 2$ with given parameters $\alpha \in \mathbb{K}$ and $\beta \in \mathbb{K}$ such that $\beta(\alpha^2 - 4) \neq 0$, is one that satisifies the following equation

$$E^M_{\alpha,\beta} : \beta v^2 = u^3 + \alpha u^2 + u$$

The curve thus defined, admits a group structure assoiated with it. For our purposes, on a finite field $\mathbb{F}_p,\ p > 2$, this curve defines a group $(E^M_{\alpha,\beta}(\mathbb{F}_p), \oplus^M_p)$ where the superscript M refers to "Montgomery". Here too, one could define the binary operation $\oplus^M_p$ from basic principles using the chord and tangent method presented earlier for Weierstrass curves in short form. However, we show shortly that every Montgomery curve is equivalent to another one in short Weierstrass form. As is the case for Twisted Edward curves, this equivalence implicitly defines a corresponding group structure associated with it. The underlying set of the group is defined as

$$E^M_{\alpha,\beta}(\mathbb{F}_p) \equiv \{(u,v) \in \mathbb{F}^2_p \mid \beta v^2 \equiv u^3 + \alpha u^2 + u \pmod p\} \cup \{\mathcal{O}^M\}$$

Note that it contains a point at infinity denoted by $\mathcal{O}^M$. The need for such a point will be addressed when we discuss the equivalence betwen curve representations.

Every Montgomery curve is equivalent to a short Weierstrass one Starting with a Montgomery curve

$$E^M_{\alpha,\beta}(\mathbb{F}_p) \equiv \{(u,v) \in \mathbb{F}^2_p \mid \beta v^2 \equiv u^3 + \alpha u^2 + u \pmod p\} \cup \{\mathcal{O}^M\},$$

where $\beta(\alpha^2 - 4) \neq 0 \pmod p$ (we will justify this constraint shortly), our objective is to transform it into a short-form Weierstrass curve

$$E^W_{A,B}(\mathbb{F}_p) \equiv \{(x,y) \in \mathbb{F}_p^2 \mid y^2 \equiv x^3 + Ax + b \pmod p\} \cup \{\mathcal{O}^W\}$$

where the superscript W refers to Weierstrass. Moreover, it must be that $p \notin \{2,3\}$ for the short Weierstrass form to hold, and that $4A^3 + 27B^2 \neq 0 \pmod p$ for it to be non-singular.

The sought after transformation must map every point (u,v) on $E^M_{\alpha,\beta}(\mathbb{F}_p)$ to a point (x,y) on $E^W_{A,B}(\mathbb{F}_p)$. Let's first exclude the point at infinity of $E^M_{\alpha,\beta}(\mathbb{F}_p)$ and focus on the other points. Let's substitute u with $(x\beta - \frac{\alpha}{3})$ and v with $y\beta$. This yields

$$\beta(y^2\beta^2) \equiv [(x\beta - \tfrac{\alpha}{3})]^3 + \alpha[(x\beta - \tfrac{\alpha}{3})]^2 + (x\beta - \tfrac{\alpha}{3}) \pmod p$$

$$\iff 27y^2\beta^3 \equiv (3x\beta - \alpha)^3 + 3\alpha(3x\beta - \alpha)^2 + 9(3x\beta - \alpha) \pmod p$$

$$\equiv 27x^3\beta^3 - 27x^2\beta^2\alpha + 9x\beta\alpha^2 - \alpha^3 + 27x^2\alpha\beta^2 + 3\alpha^3 - 18\alpha^2x\beta + 27x\beta - 9\alpha \pmod p$$

In order to make the coefficient of y^2 equal to 1 (as mandated by the short Weierstrass form), it must be that $\beta \neq 0 \pmod p$ so that we can multiply both sides of the equation by the modular inverse of $27\beta^3$. We get

$$y^2 \equiv x^3 - \tfrac{9x\beta\alpha^2}{27\beta^3} + \tfrac{2\alpha^3}{27\beta^3} + \tfrac{x}{\beta^2} - \tfrac{9\alpha}{27\beta^3} \pmod p$$

$$\iff y^2 \equiv x^3 + (\tfrac{3-\alpha^2}{3\beta^2})x + (\tfrac{2\alpha^3 - 9\alpha}{27\beta^3}) \pmod p$$

We recognize a short Weierstrass form with $A \equiv \frac{3-\alpha^2}{3\beta^2}$ and $B \equiv \frac{2\alpha^3-9\alpha}{27\beta^3}$. For it to be valid, we still need to ensure that $\Delta \equiv 4A^3 + 27B^2 \neq 0 \pmod p$. This means that

$$4(\tfrac{3-\alpha^2}{3\beta^2})^3 + 27(\tfrac{2\alpha^3-9\alpha}{27\beta^3})^2 \neq 0 \pmod p$$

$$\Rightarrow 4(3-\alpha^2)^3 + (2\alpha^3 - 9\alpha)^2 \neq 0 \pmod p \Rightarrow \alpha^2 \neq 4 \pmod p$$

The constraints $\beta \neq 0 \pmod p$ and $\alpha^2 \neq 4 \pmod p$ can be combined into a single one given by $\beta(\alpha^2 - 4) \neq 0 \pmod p$. This explains its inclusion earlier when we defined the Montgomery form.

The above derivation mapped every point of the Montgomery curve to a point on the short-Weierstrass curve. Note that the point at inifnity $\mathcal{O}^W$ of the short-Weierstrass form was not attained by the previous transformation. As a result, we define a point at infinity $\mathcal{O}^M$ on the Montgomery curve and map it to $\mathcal{O}^W$. Consequently, we get the following injective map:

$$\phi : E^M_{\alpha,\beta}(\mathbb{F}_p) \to E^W_{A,B}(\mathbb{F}_p)$$

$$(u,v) \to (x,y) \equiv \phi(u,v) = (\tfrac{u}{\beta} + \tfrac{\alpha}{3\beta}, \tfrac{v}{\beta}), \text{ if } (u,v) \neq \mathcal{O}^M$$

$$\mathcal{O}^M \to \phi(\mathcal{O}^M) = \mathcal{O}^W$$

In order to show that this map is a bijection, we must demonstrate that it has an inverse. We claim that given $\alpha, \beta \in \mathbb{F}_p$ such that $\beta(\alpha^2 - 4) \neq 0 \pmod{p}$, the short Weierstrass form $E^W_{A,B} : y^2 \equiv x^3 + (\frac{3-\alpha^2}{3\beta^2})x + \frac{2\alpha^3-9\alpha}{27\beta^3} \pmod{p}$ can be transformed into the Montgomery curve $E^M_{\alpha,\beta} : \beta v^2 \equiv u^3 + \alpha u^2 + u \pmod{p}$. Here, parameters A and B are respectively given by $(\frac{3-\alpha^2}{3\beta^2})$ and $(\frac{2\alpha^3-9\alpha}{27\beta^3})$.

This can be readily verified by substituting x with $\frac{u}{\beta} + \frac{\alpha}{3\beta}$ and y with $\frac{v}{\beta}$. This substitution shows that every point on the given short Weierstrass form is mapped to a point on the Montgomery curve. The only point left out is $\mathcal{O}^W$, which we then map to $\mathcal{O}^M$. As a result, we get the following inverse transformation:

$$\phi^{-1} : E^W_{A,B}(\mathbb{F}_p) \rightarrow E^M_{\alpha,\beta}(\mathbb{F}_p)$$

$$(x, y) \rightarrow (u, v) \equiv \phi^{-1}(x, y) = (x\beta - \tfrac{\alpha}{3},\ y\beta), \text{ if } (x, y) \neq \mathcal{O}^W$$

$$\mathcal{O}^W \rightarrow \phi^{-1}(\mathcal{O}^W) = \mathcal{O}^M$$

Note that with the exception of $\mathcal{O}^W$ and $\mathcal{O}^M$, the map ϕ (and its inverse) have their two components expressed as a rational fraction in $\mathbb{F}_p$. Such transformations are known as birational maps. As a result, the bijection between a Montgomery form and its associated short Weierstrass form is also referred to as a birational equivalence. One important observation is that any Montgomery form can be transformed into a short Weierstrass curve. However, the reverse is not always possible. We will not define the constraints that must be imposed on a short Weierstrass curve to admit a Montgomery counterpart. Suffice it to say that the specific values of A and B previously used satisfy the required constraints.

Equivalence of (certain) Twisted Edward and (certain) Montgomery curves
Starting with a Twisted Edward curve

$$E^E_{a,d}(\mathbb{F}_p) \equiv \{(x, y) \in \mathbb{F}_p^2 \mid ax^2 + y^2 \equiv 1 + dx^2y^2 \pmod{p}\},$$

where $ad(a - d) \neq 0 \pmod{p}$, and requiring additionally that a be a quadratic residue over $\mathbb{F}_p$ and d a quadratic non-residue (we will justify these constraints shortly), our objective is to transform it into a Montgomery curve given by

$$E^M_{\alpha,\beta}(\mathbb{F}_p) \equiv \{(u, v) \in \mathbb{F}_p^2 \mid \beta v^2 \equiv u^3 + \alpha u^2 + u \pmod{p}\} \cup \{\mathcal{O}^M\},$$

where $\beta(\alpha^2 - 4) \neq 0 \pmod{p}$, i.e., $\beta \neq 0 \pmod{p}$ and $\alpha \pmod{p} \notin \{-2, 2\}$.

The transformation must map every point (x, y) on $E^E_{a,d}(\mathbb{F}_p)$ to a point (u, v) on $E^M_{\alpha,\beta}(\mathbb{F}_p)$. To do so, let's substitute x with $\frac{u}{v}$ and y with $\frac{u-1}{u+1}$. This substitution attains every point on the Montgomery curve except for the following points:

1. The point at infinity $\mathcal{O}^M$.
2. Points with $v \equiv 0 \pmod{p}$: The equation of a Montgomery elliptic curve would then dictate that $u(u^2 + \alpha u + 1) \equiv 0 \pmod{p}$. This in turn means that $u \equiv 0$

$(\text{mod } p)$ or $(u^2 + \alpha u + 1) \equiv 0 \ (\text{mod } p)$. To minimize the number of points on the Montgomery curve that are not attained, we require that the quadratic equation $u^2 + \alpha u + 1$ admit no solution in $\mathbb{F}_p$. In order to ensure that, the discriminant $\alpha^2 - 4$ must be a quadratic non-residue over $\mathbb{F}_p$. Consequently, the only point to exclude from the Montgomery curve in this case is $(u, v) \equiv (0, 0)$.

3. Points with $u \equiv -1 \ (\text{mod } p)$: The equation of a Montgomery elliptic curve would then dictate that $\beta v^2 \equiv \alpha - 2 \ (\text{mod } p)$. Since $\beta \neq 0 \ (\text{mod } p)$, this becomes $v^2 \equiv \frac{\alpha-2}{\beta} \ (\text{mod } p)$. To minimize the number of points on the Montgomery curve that are not attained, we require that $\frac{\alpha-2}{\beta}$ also be a quadratic non-residue over $\mathbb{F}_p$.

The change of variable dictates that if $u \neq 0 \ (\text{mod } p)$, then $x \neq 0 \ (\text{mod } p)$. Consequently, we must exclude the points on the Twisted Edward curve that have $x \equiv 0 \ (\text{mod } p)$. One can readily verify that these are the points $(x, y) \equiv (0, 1)$ and $(x, y) \equiv (0, -1)$. So let's apply the variable substitution, keeping these two points out for now. We will treat them separately in a moment.

$$ax^2 + y^2 \equiv 1 + dx^2y^2 \ (\text{mod } p) \iff a(\tfrac{u}{v})^2 + (\tfrac{u-1}{u+1})^2 \equiv 1 + d(\tfrac{u}{v})^2(\tfrac{u-1}{u+1})^2 \ (\text{mod } p)$$

And since we excluded the cases $v \equiv 0 \ (\text{mod } p)$ and $u \equiv -1 \ (\text{mod } p)$, we get

$$\iff au^2(u+1)^2 + v^2(u-1)^2 \equiv v^2(u+1)^2 + du^2(u-1)^2 \ (\text{mod } p)$$

$$\iff (a-d)u^4 + (2a+2d)u^3 + (a-d)u^2 - 4uv^2 \equiv 0 \ (\text{mod } p)$$

Since $u \neq 0 \ (\text{mod } p)$, we can multiply both sides of the equation by the modular inverse of u and get

$$(a-d)u^3 + (2a+2d)u^2 + (a-d)u \equiv 4v^2 \ (\text{mod } p)$$

To ensure that the coefficient of u^3 is equal to 1 as mandated by the Montgomery form, we must have $a \neq d \ (\text{mod } p)$ so that we can multiply both sides of the equation by the modular inverse of $(a - d)$. In this case, we obtain

$$(\tfrac{4}{a-d})v^2 \equiv u^3 + (\tfrac{2a+2d}{a-d})u^2 + u \ (\text{mod } p)$$

We recognize the Montgomery elliptic curve form with $\alpha \equiv \frac{2a+2d}{a-d}$ and $\beta \equiv \frac{4}{a-d}$. However, we must still make sure that $\beta(\alpha^2 - 4) \neq 0 \ (\text{mod } p)$ as mandated by the definition of a Montgomery form. Clearly, $\beta \equiv \frac{4}{a-d} \neq 0 \ (\text{mod } p)$. We only need to make sure that $\alpha^2 - 4 \neq 0 \ (\text{mod } p)$. This translates to $(\frac{2a+2d}{a-d})^2 - 4 \neq 0 \ (\text{mod } p)$, which implies that $ad \neq 0 \ (\text{mod } p)$.

To sum-up, the variable substitution that we introduced defines a map from $E^E_{a,d}(\mathbb{F}_p) - \{(0, 1), (0, -1)\}$ to $E^M_{\alpha,\beta}(\mathbb{F}_p) - \{(0, 0), \ \mathcal{O}^M\}$ given by $(u, v) \equiv \psi(x, y) = (\frac{1+y}{1-y}, \ \frac{1+y}{x(1-y)})$. The constraints that need to be observed are the following:

1. $\frac{\alpha-2}{\beta}$ must be a quadratic non-residue over $\mathbb{F}_p$. In terms of the Twisted Edward

curve parameters a and d, this translates to $(\frac{2a+2d}{a-d} - 2)/\frac{4}{a-d} = d$ being a quadratic non-residue.

2. $(\alpha^2 - 4)$ must be a quadratic non-residue over $\mathbb{F}_p$. This means that $[(\frac{2a+2d}{a-d})^2 - 4]$ is a quadratic non-residue. Consequently, $\frac{16ad}{(a-d)^2}$ must be a quadratic non-residue. Since 16 and $(a-d)^2$ are both quadratic residues, it must be that ad is a quadratic non-residue. Now note that if both a and d were squares, then so will ad. Moreover, if a and d were both non-squares, there could still be a possibility that ad is a square (as an example, take the elements 3 and 5 over $\mathbb{F}_7$ which are both quadratic non-residues but their product is a square). As a result, we require that if a is a square then d be not and vice-versa. Since we saw that d is a quadratic non-residue, then we require that a be a quadratic residue.

3. $a \neq d \pmod p$ and $ad \neq 0 \pmod p$. We can combine both constraints into a single one $ad(a-d) \neq 0 \pmod p$.

Finally, note that we left out two points on the Twisted Edward curve, namely $(x, y) \equiv (0, 1)$ and $(x, y) \equiv (0, -1)$. Observe however, that on the Montgomery elliptic curve, we also have two points that were not covered, namey the point $(u, v) \equiv (0, 0)$ and the point at infinity $\mathcal{O}^M$. We thus define the following injective transformation:

$$\psi : E^E_{a,d}(\mathbb{F}_p) \rightarrow E^M_{\alpha,\beta}(\mathbb{F}_p)$$

$$(x, y) \rightarrow (u, v) \equiv \psi(x, y) = (\tfrac{1+y}{1-y},\ \tfrac{1+y}{x(1-y)}), \text{ if } (x, y) \notin \{(0, 1), (0, -1)\}$$

$$(0, -1) \rightarrow \psi(0, -1) = (0, 0)$$

$$(0, 1) \rightarrow \psi(0, 1) = \mathcal{O}^M$$

Conversely, starting with the Montgomery curve

$$E^M_{\alpha,\beta}(\mathbb{F}_p) \equiv \{(u, v) \in \mathbb{F}^2_p \mid \beta v^2 \equiv u^3 + \alpha u^2 + u \pmod p\} \cup \{\mathcal{O}^M\},$$

where $\beta(\alpha^2 - 4) \neq 0 \pmod p$, $(\frac{\alpha-2}{\beta})$ and $(\alpha^2 - 4)$ are both quadratic non-residues over $\mathbb{F}_p$, we can show that it can be transformed into a Twisted Edward curve of the form

$$E^E_{a,d}(\mathbb{F}_p) \equiv \{(x, y) \in \mathbb{F}^2_p \mid ax^2 + y^2 \equiv 1 + dx^2y^2 \pmod p\},$$

where $ad(a-d) \neq 0 \pmod p$.

To do so, we make use of the inverse of the previous substitution. We substitute u with $(\frac{1+y}{1-y})$ and v with $(\frac{1+y}{x(1-y)})$. Clearly, these are defined for all points (x, y) as long as $x \neq 0 \pmod p$ and $y \neq 1 \pmod p$. But note that $x \equiv 0 \pmod p$ corresponds to $y^2 \equiv 1 \pmod p$. In other words, points $(0, 1)$ and $(0, -1)$ of the Twisted Edward curve cannot be attained by the map defined by this substitution. Note also that the change of variable dictates that if $x \neq 0 \pmod p$, $y \neq -1 \pmod p$, and $y \neq 1 \pmod p$, then $u, v \neq 0 \pmod p$:

1. Case $u \equiv 0 \pmod{p}$: Since $\beta \neq 0 \pmod{p}$, the Montgomery form dictates that $v \equiv 0 \pmod{p}$. As a result the point $(0, 0)$ of the Montgomery curve must be excluded.

2. Case $v \equiv 0 \pmod{p}$: The Montgomery form dictates that $u(u^2 + \alpha u + 1) \equiv 0 \pmod{p}$. And so $u \equiv 0 \pmod{p}$ or $u^2 + \alpha u + 1 \equiv 0 \pmod{p}$. Since we assumed that $(\alpha^2 - 4)$ is a quadratic non-residue, the discrimant of $(u^2 + \alpha u + 1)$ cannot be a square. Hence, this quadratic has no roots over $\mathbb{F}_p$.

Consequently, the only two points of the Montgomery curve that are not covered by this substitution are the point at infinity $\mathcal{O}^M$ and the point $(0, 0)$. We will deal with them separately. After applying the change of variable, we get

$$\beta(\tfrac{1+y}{x(1-y)})^2 \equiv (\tfrac{1+y}{1-y})^3 + \alpha(\tfrac{1+y}{1-y})^2 + (\tfrac{1+y}{1-y}) \pmod{p}$$

After simplication, we obtain

$$\beta(1 + y)(1 - y^2) \equiv [(2 + \alpha)x^2 + (2 - \alpha)x^2y^2](1 + y) \pmod{p}$$

Since $y \neq -1 \pmod{p}$ and $\beta \neq 0 \pmod{p}$, we simplify further and obtain

$$(\tfrac{2+\alpha}{\beta})x^2 + y^2 \equiv 1 - (\tfrac{2-\alpha}{\beta})x^2y^2$$

We recognize this as a Twisted Edward form with $a = \frac{2+\alpha}{\beta}$ and $d = \frac{\alpha-2}{\beta}$. This is a valid representation because

1. $ad = (\frac{2+\alpha}{\beta})(\frac{\alpha-2}{\beta}) = \frac{\alpha^2-4}{\beta^2} \neq 0 \pmod{p}$ (since $(\alpha^2 - 4) \neq 0 \pmod{p}$ for a Montgomery curve).

2. $a - d = \frac{4}{\beta} \neq 0 \pmod{p}$

We then get the following inverse transformation:

$$\psi^{-1} : E^M_{\alpha,\beta}(\mathbb{F}_p) \to E^E_{a,d}(\mathbb{F}_p)$$

$$(u, v) \to (x, y) \equiv \psi^{-1}(u, v) = (\tfrac{u}{v},\ \tfrac{u-1}{u+1}), \text{ if } (u, v) \neq (0, 0)$$

$$(0, 0) \to \psi^{-1}(0, 0) = (0, -1)$$

$$\mathcal{O}^M \to \psi^{-1}(\mathcal{O}^M) = (0, 1)$$

The maps ψ and ψ^{-1} demonstrate the existence of a birational equivalence between the following two sets:

1. Twisted Edward curves with $ad(a - d) \neq 0 \pmod{p}$, a a quadratic residue and d a quadratic non-residue over $\mathbb{F}_p$

2. Montgomery curves with $\beta(\alpha^2 - 4) \neq 0 \pmod{p}$, $(\frac{2+\alpha}{\beta})$ a quadratic residue and $(\frac{\alpha-2}{\beta})$ a quadratic non-residue over $\mathbb{F}_p$.

Finally, note that one could transform such a Twisted Edward curve into a short form Weierstrass curve by applying the composition map $\phi \circ \psi$. Readers interested in learning more about Twisted Edward curves can consult [2].

Monero's curve The elliptic curve cryptography used by Monero [8] relies on a particular Twisted Edward curve known as $ed25519$ [3]. It has the following attributes:

- p is the prime number given by $2^{255} - 19$, explaining the suffix in the curve's name. It defines the underlying finite field $\mathbb{F}_p$. In hex notation using 256 bit-long representation, it is given by

$$p = \text{7FFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFED}$$

- $a \equiv -1 \pmod{p}$. In standard hex notation, it is given by

$$a = \text{7FFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFEC}$$

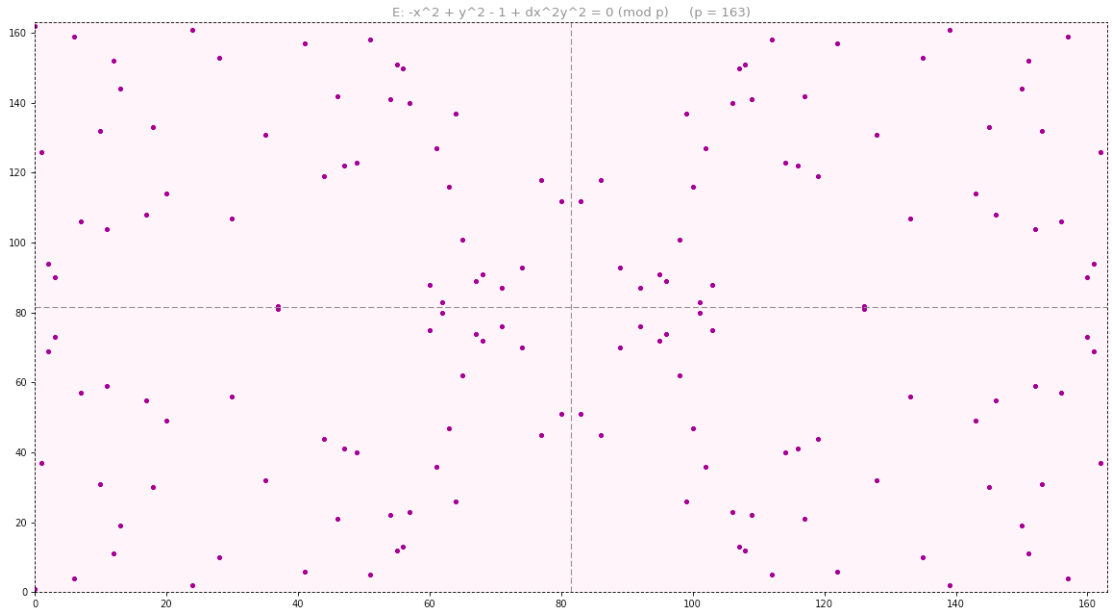
- $d \equiv -\frac{121665}{121666} \pmod{p}$. In standard hex notation, it is given by

$$d = \text{52036CEE 2B6FFE73 8CC74079 7779E898 00700A4D 4141D8AB 75EB4DCA 135978A3}$$

  Since $p \neq 2$ and $ad(a-d) \neq 0 \pmod{p}$, this curve qualifies as a Twisted Edward curve. Moreover, a is a quadratic residue over $\mathbb{F}_p$ while d is not. Consequently, this curve admits a birationally equivalent Montgomery form known as Curve25519. The underlying set of the group associated with this Twisted Edward curve is given by

$$E^E_{a,d}(\mathbb{F}_p) \equiv \{(x,y) \in \mathbb{F}_p^2 \mid -x^2 + y^2 \equiv 1 - (\tfrac{121665}{121666})x^2y^2 \pmod{p}\}$$

  Here is a euclidean representation of this curve when $p = 163$ *(it is not feasible to show it for the value of* $p = 2^{255} - 19$*).*



- The base point G has abscissa and ordinate given by

$$x_G \equiv 15112221349535400772501151409588531511454012693041857206046113283949847762202 \pmod{p}$$

$$y_G \equiv \frac{4}{5} \pmod{p} \equiv 463168356949264781694283940034751631413079938662562256157830336031652518559 60 \pmod{p}$$

which in standard hex notation are given by:

$$x_G = 216936D3\ CD6E53FE\ C0A4E231\ FDD6DC5C\ 692CC760\ 9525A7B2\ C9562D60\ 8F25D51A$$

$$y_G = 66666666\ 66666666\ 66666666\ 66666666\ 66666666\ 66666666\ 66666666\ 66666658$$

Monero's public-key cryptography is hence conducted on the subgroup whose underlying set is $\{G\}$.

- The order of G is chosen to be a prime number equal to

$$n \equiv 2^{252} + 27742317777372353535851937790883648493 \pmod{p}$$

Which in standard hex notation is given by

$$n = 10000000\ 00000000\ 00000000\ 00000000\ 14DEF9DE\ A2F79CD6\ 5812631A\ 5CF5D3ED$$

- Recall that n denotes the order of G, and must divide $\#E^E_{a,d}(\mathbb{F}_p)$ i.e., the order of

$E^E_{a,d}(\mathbb{F}_p)$. The cofactor h is equal to $\frac{\#E^E_{a,d}(\mathbb{F}_p)}{n}$ which in this case evaluates to 8. It is represented in standard hex notation by

$h =$ 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000008

References

[1] behindtext. why did bitcoin choose secp256k1 over secp256r1? https://bitcointalk.org/index.php?topic=151120.0.

[2] D. J. Bernstein, Peter Birkner, Marc Joye, Tanja Lange, and Christiane Peters. Twisted edwards curves. *AFRICACRYPT*, pages 389–405, 2008.

[3] Daniel J. Bernstein, Niels Duif, Tanja Lange, Peter Schwabe, and Bo Yin Yang. High-speed high-security signatures. *Cryptographic Hardware and Embedded Systems*, pages 124–142, 2011.

[4] Anne-Sophie Charest. Pollard's p-1 and lenstra's factoring algorithms. http://www.math.mcgill.ca/darmon/courses/05-06/usra/charest.pdf, October 2005.

[5] Hal Finney. secp256k1. https://bitcointalk.org/index.php?topic=2699.0.

[6] ImperialViolet. What a difference a prime makes. https://www.imperialviolet.org/2010/12/21/eccspeed.html, December 2010.

[7] Certicom Research. Sec 2: Recommended elliptic curve domain parameters. *Standards for Efficient Cryptography*, 2010.

[8] N. Van Saberhagen. Cryptonote 2.0. https://cryptonote.org/whitepaper.pdf, 2013.

[9] R. Schoof. Elliptic curves over finite fields and the computation of square roots mod p. *Mathematics of computation*, 44(170):483–494, 1985.

[10] J. Silverman. *The Arithmetic of Elliptic Curves*. Springer, second edition, 2009.

[11] Nick Sullivan. How the nsa (may have) put a backdoor in rsa's cryptography: A technical primer. https://blog.cloudflare.com/how-the-nsa-may-have-put-a-backdoor-in-rsas-cryptography-a-technical-primer/, January 2014.