Elliptic curve cryptography

Chloe Martindale

These lecture notes are for the short course on isogeny-based cryptography given at the SLMath summer school at IBM Research Zürich in June 2024.

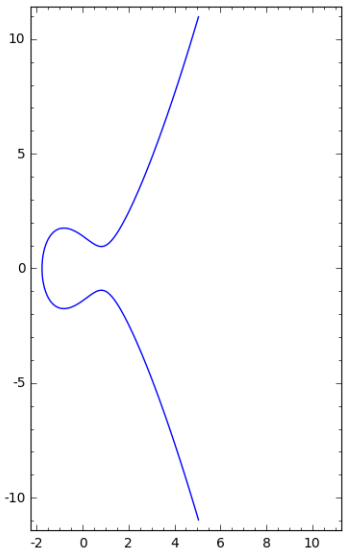
1 Introduction to elliptic curves

Definition 1. Let K be a field with characteristic different from 2 or 3 (for $K = \mathbb{F}_{p^r}$ this means that $p \neq 2$ or 3). An *elliptic curve* E defined over a field K is a curve of the form

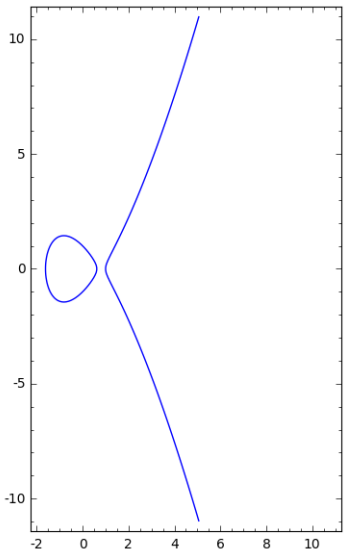
$$E : y^2 = x^3 + ax + b,$$

with $a, b \in K$ and $4a^3 + 27b^2 \neq 0$.

Examples. Here is an example with $a = -2$ and $b = 2$, i.e., the curve $y^2 = x^3 - 2x + 2$:

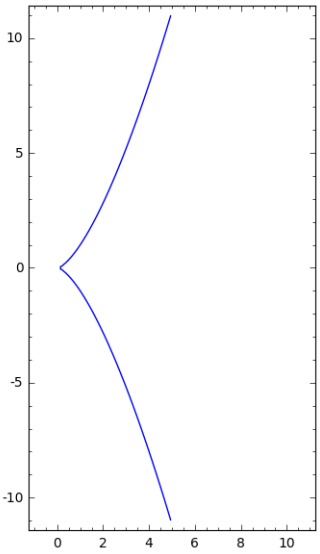


Here is an example with $a = -2$ and $b = 1$, i.e., the curve $y^2 = x^3 - 2x + 1$:

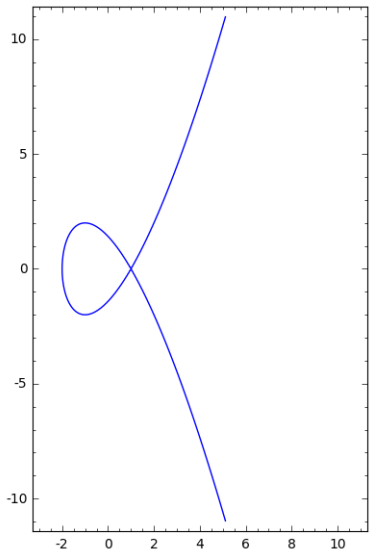


To see why we have excluded a and b such that $4a^3 + 27b^2 = 0$, consider the following non-examples of elliptic curves:

- $a = b = 0$, the curve $y^2 = x^3$:



- $a = -3$, $b = 2$, the curve $y^2 = x^3 - 3x + 2$:



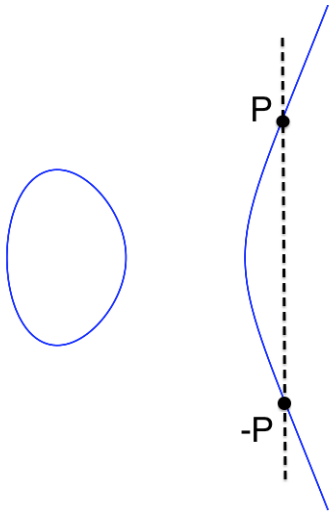
So we include the condition $4a^3+27b^2 = 0$ to avoid curves with 'sharp' points or curves that cross themselves. A fundamental reason that elliptic curves are so widely studied (not just in cryptography, but this is essential for cryptography too) is that it is possible to define a group law on the set of K-rational points of an elliptic curve defined over K. The 'set of K-rational points' here refers to the solutions $(x, y) \in K \times K$ to the defining equation, together with one more point that we will discuss below. Let us first think about the case of $K = \mathbb{Q}$.

Define

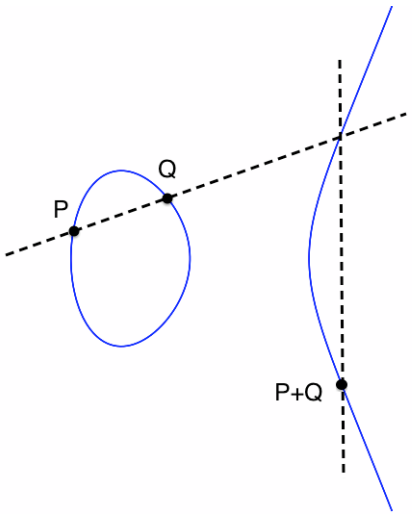
$$G = \{(x, y) \in \mathbb{Q} \times \mathbb{Q} : y^2 = x^3 + ax + b\}$$

for some $a, b \in \mathbb{Q}$. We can 'almost' make a group from G. The group law on Weierstrass curves has a nice geometric definition.

- We define the inverse of a point (x, y) to be $(x, -y)$.



- We define every vertical line to have an invisible point P_∞, 'the point at infinity', and this is the neutral element of the group.
- We define a straight line that is tangent to the curve to intersect the curve twice at that point.
- With the above conventions, every straight line passing through at least 2 points on the curve intersects the curve in exactly 3 points. We define the sum of 3 points on a straight line to be P_∞, hence addition looks like this:



We made quite a few choices in defining our group law $+$, so we need to check the group axioms to make sure that it is really a group law for $G \cup \{P_\infty\}$. Recall the group axioms:

Definition 2. We say that G is a *group* under $*$ if

(G1) For every $a, b \in G$, $a * b \in G$.

(G2) For every $a, b, c \in G$, $(a * b) * c = a * (b * c)$.

(G3) There exists $e \in G$ such that for every $a \in G$, $e * a = a * e = a$.

(G4) For every $a \in G$, there exists $b \in G$ such that $a * b = b * a = e$.

Now we can check these axioms to ensure that $G \cup \{P_\infty\}$ is a group under the operation $+$ that we defined above:

(G1) To check (G1), we need to make sure that given P and Q in $G \cup \{P_\infty\}$, $P + Q \in G \cup \{P_\infty\}$. If P or $Q = P_\infty$ this is trivial, so assume otherwise. $P + Q$ is on the curve by definition, so we only need to check that the coordinates are rational. The coordinates of $P + Q$ are rational if and only if the coordinates of $-(P + Q)$ are rational, which was the third point of intersection between the line through P and Q and the elliptic curve. Suppose that the the equation of the line through P and Q is given by $y = mx + c$. Then as P and Q have rational coordinates, m and $c \in \mathbb{Q}$. To get the third point of intersection of $y = mx + c$ with $y^2 = x^3 + ax + b$, we just plug y into E to get a cubic in x with rational coefficients, 2 roots of which (x_P and x_Q) are known to be rational, hence the third is also rational. So the x-coordinate of $-(P + Q)$ is rational, hence also the y coordinate as $y = mx + c$.

(G2) To check (G2), we need to check that given P, Q, and $R \in G \cup \{P_\infty\}$, $P + (Q + R) = (P + Q) + R$. Checking this by writing out the formulae is easy but long, so we skip it.

(G3) Axiom (G3) states that there exists a neutral element, which is P_∞ by definition.

(G4) Axiom (G4) states that every element has an inverse, which we saw already was given by reflecting about the x-axis.

Remark 1. Another way to think of P_∞ is the following. When we study elliptic curves and their associated groups, the $y^2 = x^3 + ax + b$ (with a and b in K) comes from setting $x = X/Z$ and $y = Y/Z$ in the equation

$$Y^2 Z = X^3 + aXZ^2 + bZ^3.$$

Note that every term in this equation has degree 3, so that if (X_0, Y_0, Z_0) is a solution of this equation, then (nX_0, nY_0, nZ_0) is also a solution of the equation for every n in K. For this reason, if $(nX_0, nY_0, nZ_0) = (X_0, Y_0, Z_0)$ then we say that the 2 solutions are *equivalent*. Observe that these solutions all correspond to a unique x and y! The point at infinity is

$$P_\infty = (0, 1, 0)$$

in (X, Y, Z)-coordinates, which gets 'sent to infinity' when we switch to (x, y)-coordinates. Note that this is always on the curve!

Having intuitively constructed a geometric group law elliptic curves over $\mathbb{Q}$, if we now write down the formulae for adding points, we can get a group law for elliptic curves over $\mathbb{F}_q$. So what are the formulae for adding?

Write $P = (x_P, y_P)$ and $Q = (x_Q, y_Q)$, and define $(x_R, y_R) = R = P + Q$. We want to write down a formula for x_R and for y_R. We know that P, Q, and $-R$ all lie on the straight line passing through P and Q, so we first calculate the formula of this line. The equation of this line is $y = mx + c$ where

$$m = \begin{cases} (y_Q - y_P)/(x_Q - x_P) & P \neq Q \\ (3x_P^2 + a)/(2y_P) & P = Q \end{cases}$$

and

$$c = y_P - mx_P.$$

(Recall that the gradient of a tangent line to a curve at a point P is the value of $\frac{dy}{dx}$ at P.) We plug in $y = mx + c$ with m and c as above to the equation for E and solve to find the intersection points:

$$(mx + c)^2 = x^3 + ax + b.$$

We know that the roots of this cubic are x_P, x_Q, and x_R, so

$$x^3 - (mx + c)^2 + ax + b = (x - x_P)(x - x_Q)(x - x_R).$$

Then by comparing coefficients of x^2, we see that

$$x_R = m^2 - x_P - x_Q.$$

Then we can just use the equation of the line to compute y_R:

$$y_R = -y_{-R} = -(mx_R + c).$$

With these explicit formulae, we can define, for any $a, b \in \mathbb{F}_q$ such that $4a^3 + 27b^2 \neq 0$, a group law on

$$G = \{(x, y) \in \mathbb{F}_q \times \mathbb{F}_q : y^2 = x^3 + ax + b\} \cup \{P_\infty\}$$

as

$$(x_P, y_P) + (x_Q, y_Q) = (m^2 - x_P - x_Q, -m(m^2 - x_P - x_Q) - c),$$

where

$$m = \begin{cases} (y_Q - y_P)/(x_Q - x_P) & P \neq Q \\ (3x_P^2 + a)/(2y_P) & P = Q \end{cases}$$

and

$$c = y_P - mx_P.$$

All the case distinctions with P_∞ can be avoided with some clever tricks for efficiency, which we will see below.

1.1 Efficient arithmetic with elliptic curves

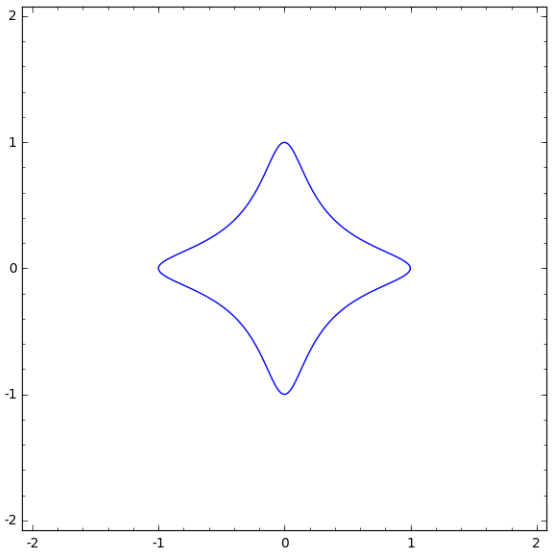
Of vital importance in real-world applications of cryptographic primitives such as CSIDH or SQISign is maximising efficiency without sacrificing security. The rich mathematical structure of elliptic curves lends itself to many ideas beyond double-and-add, some of which we study below.

1.2 Edwards curves

One form of elliptic curve that turns out to be beneficial for efficient arithmetic is an *Edwards curve*. Although the Edwards and twisted Edwards curves we will see here have an x^2y^2 term which does not appear in the Weierstrass model; more on that later.

Example. Let's try to make a group from the points on an Edwards curve. We will look first at the example

$$C : x^2 + y^2 = 1 - 30x^2y^2.$$



Note that the equation of C looks similar to the equation of a circle with a 'fudge factor', and we will see that we can construct a group law similar to that of the circle plus this 'fudge factor'. Define

$$G = \{(x, y) \in \mathbb{R} \times \mathbb{R} : x^2 + y^2 = 1 - 30x^2y^2\}.$$

Claim 1. For $(x_1, y_1), (x_2, y_2) \in G$ define

$$(x_1, y_1) \oplus (x_2, y_2) := \left(\frac{x_1y_2 + y_1x_2}{1 - 30x_1y_1x_2y_2}, \frac{y_1y_2 - x_1x_2}{1 + 30x_1y_1x_2y_2} \right).$$

Then $(G, \oplus)$ is a group with neutral element $(0, 1)$.

Proof. We first have to check that we didn't divide by zero, that is, we should check that for $(x_1, y_1), (x_2, y_2) \in G$, we never get that $1 \pm 30x_1x_2y_1y_2 = 0$. If x_1, x_2, y_1, or $y_2 = 0$ then this is clearly non-zero, so suppose that x_1, x_2, y_1, and y_2 are non-zero. Then by the curve equation, for $i = 1, 2$,

$$x_i^2 + y_i^2 + 30x_i^2y_i^2 = 1,$$

and $x_i^2 + y_i^2 > 0$ so

$$30x_i^2y_i^2 < 1,$$

hence

$$\sqrt{30}|x_iy_i| < 1.$$

Therefore

$$30|x_1x_2y_1y_2| = \sqrt{30}|x_1y_1|\sqrt{30}|x_2y_2| < 1 \cdot 1 = 1,$$

so the denominators of the operation $\oplus$ are never zero. We still need to check that it actually defines a group law, that is, that the group axioms (G1)-(G4)–recalled in the 'Elliptic curves - mathematical foundations' notes–are satisfied.

(G1) For the axiom (G1), we have to check that $(x_1, y_1) \oplus (x_2, y_2) \in G$, that is, we have to check that

$$\left(\frac{x_1y_2 + y_1x_2}{1 - 30x_1y_1x_2y_2}\right)^2 + \left(\frac{y_1y_2 - x_1x_2}{1 + 30x_1y_1x_2y_2}\right)^2$$
$$= 1 - 30\left(\frac{x_1y_2 + y_1x_2}{1 - 30x_1y_1x_2y_2}\right)^2\left(\frac{y_1y_2 - x_1x_2}{1 + 30x_1y_1x_2y_2}\right)^2,$$

which we can do just by simplification.

(G2) For the axiom (G2), we have to check that if $(x_1, y_1), (x_2, y_2)$, and $(x_3, y_3) \in G$, then

$$((x_1, y_1) \oplus (x_2, y_2)) \oplus (x_3, y_3) = (x_1, y_1) \oplus ((x_2, y_2) \oplus (x_3, y_3)),$$

which we can again doing just by plugging in the formulae and simplifying.

(G3) For the axiom (G3), we have to check that for every $(x, y) \in G$, $(x, y) \oplus (0, 1) = (0, 1) \oplus (x, y) = (x, y)$. We plug $(x_1, y_1) = (x, y)$ and $(x_2, y_2) = (0, 1)$ into our formula for $\oplus$ to get

$$(x, y) \oplus (0, 1) = \left(\frac{x \cdot 1 + 0 \cdot y}{1 - 30x \cdot y \cdot 0 \cdot 1}, \frac{y \cdot 1 - x \cdot 0}{1 + 30x \cdot y \cdot 0 \cdot 1}\right) = (x, y),$$

and similarly for $(0, 1) \oplus (x, y)$.

(G4) For the axiom (G4), we have to check that for every $(x, y) \in G$, there exists $-(x, y) \in G$ such that $(x, y) + (-(x, y)) = (0, 1)$. We claim that $-(x, y) = (-x, y)$:

$$(x, y) \oplus (-x, y) = \left(\frac{xy - xy}{1 - 30x^2y^2}, \frac{x^2 + y^2}{1 + 30x^2y^2}\right)$$
$$= (0, 1),$$

as by the curve equation $x^2 + y^2 = 1 + 30x^2y^2$.

□

Definition 3. Suppose that $d \in \mathbb{F}_q^*$ is a non-square (i.e., that for g a primitive element of $\mathbb{F}_q^*$, $d = g^k$ for k odd). Then the curve

$$C_d : x^2 + y^2 = 1 + dx^2y^2$$

is an *Edwards curve over* $\mathbb{F}_q$.

Note that the example we looked at was C_{-30} but over $\mathbb{R}$. In fact

$$(x_1, y_1) \oplus (x_2, y_2) := \left(\frac{x_1y_2 + y_1x_2}{1 + dx_1y_1x_2y_2}, \frac{y_1y_2 - x_1x_2}{1 - dx_1y_1x_2y_2} \right)$$

defines a group law of C_d just as before. Checking the group axioms is exactly the same process, but as the proof that the denominators are non-zero is different, we will write that out.

Claim 2. Suppose that $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ are on C_d, i.e. that for $i = 1, 2$

$$x_i^2 + y_i^2 = 1 + dx_i^2y_i^2.$$

Then

$$1 \pm dx_1x_2y_1y_2 \neq 0.$$

Proof. Suppose for a contradiction that

$$dx_1x_2y_1y_2 = \pm 1. \tag{1}$$

Then

$$\begin{aligned} dx_1^2y_1^2(x_2 + y_2)^2 &= dx_1^2y_1^2(x_2^2 + y_2^2 + 2x_2y_2) \\ &= dx_1^2y_1^2(1 + dx_2^2y_2^2 + 2x_2y_2) \\ &= d^2x_1^2y_1^2x_2^2y_2^2 + dx_1^2y_1^2 + 2(dx_1x_2y_1y_2)x_1y_1 \\ &= 1 + dx_1^2y_1^2 \pm 2x_1y_1 \\ &= (x_1 \pm y_1)^2, \end{aligned}$$

but as d is non-square, $dx_1^2y_1^2(x_2 + y_2)^2$ is non-square or zero, and $(x_1 \pm y_1)^2$ is square, so we must have that

$$dx_1^2y_1^2(x_2 + y_2)^2 = (x_1 \pm y_1)^2 = 0.$$

By the assumption $dx_1x_2y_1y_2 = \pm 1$ that x_1, y_1, x_2, and y_2 are non-zero, and by definition that $d = 0$, hence

$$x_2 + y_2 = 0.$$

But if (x_2, y_2) is on C_d, then $(x_2, -y_2)$ is also on C_d, hence the above argument with $y_2 = -y_2$ gives that

$$x_2 - y_2 = 0,$$

hence $x_2 = y_2 = 0$, which is a contradiction to (1). □

So we have a group under $\oplus$ made up of the $\mathbb{F}_q$-points on C_d, but how easy is arithmetic in this group? Note first of all that doubling a point is actually easier than adding 2 different points:

$$\begin{aligned} 2 \cdot (x, y) &= (x, y) \oplus (x, y) \\ &= \left(\frac{2xy}{1 + dx^2y^2}, \frac{y^2 - x^2}{1 - dx^2y^2} \right) \\ &= \left(\frac{2xy}{x^2 + y^2}, \frac{y^2 - x^2}{2 - x^2 - y^2} \right). \end{aligned}$$

These equations have lower degree than the equations for adding two different points, which means faster computation (we will see later how much faster). Still, we have to do an inversion to compute the sum of 2 points or the double of a point, but we can 'delay' this inversion. So, our aim now is to compute

$$(x_3, y_3) := (x_1, y_1) \oplus (x_2, y_2)$$

with the minimum number of inversions and multiplications. To 'delay' the inversion, we introduce new variables X_i, Y_i, Z_i and substitute $x_i = X_i/Z_i$ and $y_i = Y_i/Z_i$. Then

$$x_3 = \frac{(X_1Y_2 + X_2Y_1)Z_1Z_2}{(Z_1Z_2)^2 + dX_1X_2Y_1Y_2}$$

and

$$y_3 = \frac{Z_1Z_2(Y_1Y_2 - X_1X_2}{(Z_1Z_2)^2 - dX_1X_2Y_1Y_2}.$$

Define

$$\begin{aligned} X_3 &= Z_1Z_2(X_1Y_2 + X_2Y_1)((Z_1Z_2)^2 - dX_1X_2Y_1Y_2), \\ Y_3 &= Z_1Z_2(Y_1Y_2 - X_1X_2)((Z_1Z_2)^2 + dX_1X_2Y_1Y_2), \end{aligned}$$

and

$$Z_3 = ((Z_1Z_2)^2 - dX_1X_2Y_1Y_2)((Z_1Z_2)^2 + dX_1X_2Y_1Y_2).$$

Then $x_3 = X_3/Z_3$ and $y_3 = Y_3/Z_3$, and if we just compute X_3, Y_3, and Z_3 then we don't have to do any inversions! In fact, X_3, Y_3, and Z_3 can be computed in just 10 multiplications (M), one squaring (S), and one multiplication by (D) in the following way:

1. $A = Z_1Z_2$, $B = A^2$, $C = X_1X_2$, $D = Y_1Y_2$. (3M + 1S).
2. $E = dCD$, $F = B - E$, $G = B + E$. (1M + 1D).
3. $X_3 = AF((X_1 + Y_1)(X_2 + Y_2) - C - D)$. (3M).
4. $Y_3 = AG(D - C)$. (2M).
5. $Z_3 = FG$. (1M).

Note that in step 3 we reduced the multiplications by a clever trick:

$$X_1Y_2+X_2Y_1 = X_1Y_2+X_2Y_1+X_1X_2+Y_1Y_2-X_1X_2-Y_1Y_2 = (X_1+Y_1)(X_2+Y_2)-C-D.$$

Doubling can be done in just 4S + 3M, so here we concretely that it is much faster than adding distinct points.

You can also make scalar multiplication faster by precomputing some multiplications of P, e.g., by using that

$$15P = 8P + 4P + 2P + P.$$

1.3 Montgomery curves

Definition 4. A *Montgomery curve* is a curve of the form

$$M_{A,B} : Bv^2 = u^3 + Au^2 + u$$

for $B(A^2 - 4) \neq 0$. The group law looks very similar to the group law for Weierstrass curves:

$$(u_1, v_1) \oplus (u_2, v_2) = (Bm^2 - A - u_1 - u_2, m(u_1 - u_3) - v_1),$$

where $u_3 = Bm^2 - A - u_1 - u_2$, and

$$m = \begin{cases} (v_1 - v_2)/(u_1 - u_2) & (u_1, v_1) \neq (u_2, v_2) \\ (3u_1^2 + 2Au_1 + 1)/(2Bv_1) & (u_1, v_1) = (u_2, v_2). \end{cases}$$

The neutral element is again P_∞.

We have now mentioned a few times a 'transformation' that relates different curve shapes. We would like a way to say when 2 curves are 'the same', or at least a way to say what 'the same' means! Let's think about what this would mean in an ideal world..

Suppose that we have some map from the Edwards curve

$$C_d : x^2 + y^2 = 1 + dx^2y^2$$

to the Montgomery curve $M_{A,B}$ above given by

$$f : C_d \longrightarrow M_{A,B}.$$

It would be nice if for all points P on C_d and for all $a \in \mathbb{Z}$, we had that $f(aP) = af(P)$, as $f(P) = Q$ and $af(P) = aQ$ are points on a Montgomery curve, so then we can find a (if we can break DLP on a Montgomery curve). That is, if f satisfies this nice property of $f(aP) = af(P)$, we can somehow translate the discrete logarithm on C_d to a discrete logarithm on $M_{A,B}$. It would also be nice to be able to go the other direction, that is if there's a map

$$g : M_{A,B} \longrightarrow C_d$$

that is the inverse of f. Some other nice properties to require of f:

- $f(P+Q) = f(P) + f(Q)$
- $f((0,1)) = P_\infty$ (remember that the neutral point on C_d was $(0,1)$.

We can write down a nice map from a twisted Edwards curve to a Montgomery curve:

$$\begin{array}{ccc} ax^2 + y^2 = 1 + dx^2y^2 & \longrightarrow & Bv^2 = u^3 + Au^2 + u \\ (x,y) & \longmapsto & (u,v) = ((1+y)/(1-y), (1+y)/(x(1-y))) \\ (a,d) & \longmapsto & (A,B) = (2(a+d)/(a-d), B = 4/(a-d). \end{array}$$

In other direction we have the map:

$$\begin{array}{ccc} Bv^2 = u^3 + Au^2 + u & \longrightarrow & ax^2 + y^2 = 1 + dx^2y^2 \\ (u,v) & \longmapsto & (x,y) = (u/v, (u-1)/(u+1)) \\ (A,B) & \longmapsto & (a,d) = ((A+2)/B, (A-2)/B). \end{array}$$

This map is an 'isomorphism':

Definition 5. Let E/k and E'/k be elliptic curves. An *isomorphism* $f : E \to E'$ is a birational map that induces an isomorphism of groups $E(\overline{k}) \cong E'(\overline{k})$. In this case we say that E and E' are *isomorphic*.

By a similar transformation we can go from Montgomery to Weierstrass, but not necessarily back! All elliptic curves are Weierstrass, but not all can be written in Edwards/Montgomery form. There is a quick way to see this: Edwards curves (and hence Montgomery curves) always have a point $(1,0)$ of order 4, and there are examples of Weierstrass curves that do not. To see that $(1,0)$ is a point of order 4, recall the doubling formula for Edwards curves from last time:

$$2 \cdot (x,y) = \left(\frac{2xy}{x^2+y^2}, \frac{y^2-x^2}{2-x^2-y^2} \right).$$

Then we can easily compute

$$2 \cdot (1,0) = (0,-1)$$

and hence

$$4 \cdot (1,0) = 2 \cdot (0,-1) = (0,1).$$

(Recall that $(0,1)$ is the neutral element.)

Given an elliptic curve over k, the set of elliptic curves that are isomorphic to it is called its *isomorphism class*. Each isomorphism class of elliptic curves has an invariant called the j-invariant, which is just a number in k. For curves in Weierstrass form

$$E : y^2 = x^3 + ax + b,$$

the j-invariant can be easily computed via the formula

$$j(E) = 1728\frac{4a^3}{4a^3 + 27b^2}.$$

Some final observations on elliptic curves:

- Computations on Edwards curves are faster than on curves in Weierstrass form, so you should use Edwards curves for implementations when possible.
- Montgomery curves are typically used for isogeny-based cryptography, where we have efficient formulae for isogenies and efficient storage (see later).
- More formulae for addition/doubling on elliptic curves in various shapes are available at `hyperelliptic.org/EFD`.

2 Isogenies

Now that we have learnt about elliptic curve, and about how some elliptic curves are related via isomorphism, a natural question to ask is: are there other maps between elliptic curves? *Isogenies* are a natural generalization of isomorphisms.

Definition 6. Let E/k and E'/k be elliptic curves. An *isogeny* is a rational map $E \to E'$ that induces a surjective group homomorphism $E(\overline{k}) \to E'(\overline{k})$.

We now see some examples:

1. Let $E_{51}/\mathbb{F}_{419} : y^2 = x^3 + 51x^2 + x$ be a Montgomery elliptic curve. Then the multiplication-by-two map

$$\begin{array}{rccc} [2]: & E_{51} & \to & E_{51} \\ & P & \mapsto & P + P \end{array}$$

   is an isogeny.
   Exercise: show that [2] can be represented by the rational map

$$(x, y) \quad \mapsto \quad \left(\frac{\frac{1}{2}x^4 - 18x^3 - 163x^2 - 18x + \frac{1}{2}}{8x(x^2+9x+1)}, \frac{y(x^6+18x^5+5x^4-5x^2-18x-1)}{(8x(x^2+9x+1))^2}\right).$$

2. Let E_{51} be as above and let $E_9/\mathbb{F}_{419} : y^2 + x^3 + 9x^2 + x$. Then the map

$$\begin{array}{rccc} f: & E_{51} & \to & E_9 \\ & (x, y) & \mapsto & \left(\frac{x^3-183x^2+73x+30}{(x+118)^2}, \frac{y(x^3-65x^2-104x+174)}{(x+118)^3}\right) \end{array}$$

   is an isogeny. This isogeny has degree 3 and its kernel is given by

$$\ker(f) = \{\mathcal{O}, (-118, 51), (-118, -51)\}.$$

This kernel is a subgroup of $E_{51}(\mathbb{F}_{419})$ and is cyclic and generated by $(-118, 51)$, a point of order 3. There are other points of order 3 on E_{51} defined over an extension field, for example $Q = (210, \sqrt{380}) \in E_{\mathbb{F}_{419^2}}$. If we 'push' Q through f, we get a point $f(Q) \in E_9$, which still has order 3. This point has a special purpose: there exists another isogeny $g : E_9 \to E_{51}$ whose kernel is generated by $f(Q)$, and the composition $g \circ f = [3]$, the multiplication-by-3 map on E_{51}. This is not a coincidence: we can always construct such a map, it is called the *dual map* (formally defined below).

You will notice that in the second example (and the first), the description of the kernel is much simpler than the expression in terms of rational maps. In fact it can be even simpler if you give only the generators of the kernel. The wonderful thing is: this is enough - an isogeny is uniquely determined by its kernel (we will state this formally later) - and we even have Vélu's formulae to recover the rational maps from the kernel.

Vélu observed in his seminar paper in 1973 that, if

$$\begin{array}{rccc} f: & E & \to & E' \\ & (x,y) & \mapsto & (X,Y) \end{array}$$

is an isogeny, and we denote by (x_P, y_P) the affine coordinates of a point P on E, then

$$X_{f(P)} = x_P + \sum_{Q \in \ker(f) - \{\mathcal{O}\}} (x_{P+Q} - x_Q)$$

and

$$Y_{f(P)} = y_P + \sum_{Q \in \ker(f) - \{\mathcal{O}\}} (y_{P+Q} - y_Q).$$

He also gives an explicit derivation of the equations when E is in Weierstrass form, as well as the equation of the codomain E'; the paper is (short and) available in English for free at `https://aghitza.org/publications/translation-velu.pdf`.

Many of the points that came up in the examples can be captured in the following formal definitions and theorems:

Definition 7. Let $E, E'/\mathbb{F}_{p^r}$ be elliptic curves and let $\ell \in \mathbb{Z}_{>0}$ such that p and ℓ are coprime. An *ℓ-isogeny* $f : E \to E'$ is an isogeny with $\#\ker(f) = \ell$.

Definition 8. Let $E/\mathbb{F}_{p^r}$ be an elliptic curve and let $\ell \in \mathbb{Z}_{>0}$ such that p and ℓ are coprime. Let $f : E \to E'$ be an ℓ-isogeny. Then there exists a unique (up to isomorphism) isogeny $f^\vee : E' \to E$ such that $f^\vee \circ f = [\ell]$. This is called the *dual isogeny*. (See the example above for how to compute such an isogeny!)

Theorem. *(Isogenies are uniquely defined by their kernels). There is a one-to-one correspondence from finite subgroups of an elliptic curve to separable isogenies from said curves, up to post-composition with isomorphisms.*

Theorem. *Let E/k be an elliptic curve, and let ℓ be a prime coprime to the characteristic of k. Then, over $\overline{k}$, there are $\ell + 1$ non-isomorphic ℓ-isogenies from E.*

This theorem follows from the fact that isogenies are uniquely defined by their kernels (up to isomorphism), and that the *ℓ-torsion subgroup of E*

$$E[\ell] = \{P \in E(\overline{k}) : [\ell]P = \mathcal{O}\} \cong \mathbb{Z}/\ell\mathbb{Z} \times \mathbb{Z}/\ell\mathbb{Z}.$$

Definition 9. Let S be a set of primes and let k be a finite field. The *supersingular S-isogeny graph over k* as as

- Vertices: isomorphism classes of supersingular elliptic curves. (Can be k or $\overline{k}$-isomorphism, depending on the setup.)
- Edges: for any $\ell \in S$, an edge $E - E'$ represents an ℓ-isogeny $E \to E'$ and its dual.

3 Introduction to quaternion algebras

In the slides, we gave a high-level overview of SQISign, but before we can say how the verification arrow is computed in zero-knowledge we need to cover a couple of facts about endomorphism rings of supersingular elliptic curves: Now we will consider all the endomorphisms over the algebraic closure, not just those over the base field.

Theorem. *Let $E/\mathbb{F}_{p^r}$ be a supersingular elliptic curve. Then there exists a supersingular elliptic curve $E'/\mathbb{F}_{p^2}$ that is $\overline{\mathbb{F}}_p$-isomorphism to E.*

This theorem means that the vertices of the supersingular isogeny graph for $\mathbb{F}_{p^r}$, for any $r \geq 2$, can all be represented by j-invariants in $\mathbb{F}_{p^2}$. Even better, this graph is actually connected for $S = \ell$, for any prime $\ell \neq p$, so contains *every* supersingular j-invariant over $\mathbb{F}_{p^2}$:

Theorem. *Let $\ell \neq p$ be prime. Then the supersingular ℓ-isogeny graph over $\mathbb{F}_{p^2}$ is connected.*

This graph is even $\ell + 1$-regular, by the theorem above that says that over the algebraic closure there are always $\ell + 1$ isogenies of degree ℓ from any given elliptic curve. We do run into trouble when a vertex has automorphisms of degree ℓ, but this happens at only one or two vertices in the whole graph.

Now let's relate all this to quaternion algebras:

Theorem. *Let E/k be an elliptic curve. Then $End(E)$ is one of:*

- $\mathbb{Z}$,
- *an order in a quadratic number ring, or*
- *a maximal order in a quaternion algebra.*

In fact, it turns out that an elliptic curve is supersingular if and only if its endomorphism ring is a maximal order in a quaternion algebra (a concept we will define below), so this is an alternative characterisation of supersingular curves: and fundamental to SQISign.

Definition 10. A *quaternion algebra* B is a rank 4 $\mathbb{Q}$-algebra

$$B = \mathbb{Q} + i\mathbb{Q} + j\mathbb{Q} + k\mathbb{Q}$$

such that there exist $\alpha, \beta \in \mathbb{Z}_{>0}$ for which $i^2 = -\alpha$, $j^2 = -\beta$, $k = ij = -ji$.

Note that in particular B is *not commutative*. We have natural concepts of conjugation, norm, and trace which are similar to the complex numbers:

Definition 11. Let $x = a + ib + jc + kd \in B$. The *conjugate* of x is

$$\overline{x} = a - ib - jc - kd,$$

the *reduced norm* of x is

$$\mathrm{nrd}(x) = x\overline{x}$$

and the *reduced trace* of x is

$$\mathrm{trd}(x) = x + \overline{x}.$$

If you are familiar with number fields, then the following concept will also seem familiar:

Definition 12. Let B be a quaternion algebra. An *order* of B is a rank 4 $\mathbb{Z}$-lattice that is a subring of B.

One consequence of an element x lying in an order is that $\mathrm{nrd}(x) \in \mathbb{Z}$. A simple example of an order is the following:

$$\mathcal{O} = \mathbb{Z} + i\mathbb{Z} + j\mathbb{Z} + k\mathbb{Z}.$$

Definition 13. Let B be a quaternion algebra. A *maximal order* in B is an order that is not contained in any other order.

Unlike in number rings, maximal orders in quaternion algebras are not unique! Two maximal orders $\mathcal{O}$ and $\mathcal{O}'$ are isomorphic when there exists $\alpha \in B$ such that $\mathcal{O}' = \alpha\mathcal{O}\alpha^{-1}$, but even up to isomorphism there can be many maximal orders in a given B.

We will see shortly that maximal quaternion orders are going to correspond to supersingular elliptic curves via their endomorphism rings. But what do the isogenies correspond to?

Definition 14. Let $\mathcal{O}$ be a maximal order in a quaternion algebra B. A *left-ideal* (resp. right-ideal) of $\mathcal{O}$ is an additive subgroup of $\mathcal{O}$ satisfying

$$\mathcal{O}I \subseteq I$$

(resp. $I\mathcal{O} \subseteq I$).

Definition 15. Let I be a left- or right-ideal in B. Then the *left-order* (resp. right-order) of I is given by

$$\mathcal{O}_\ell(I) := \{x \in B : xI \subseteq I\}$$

(resp. $\mathcal{O}_r(I) := \{x \in B : Ix \subseteq I\}$).

Definition 16. Let I be a left- or right-ideal in B. The *norm* of I is

$$N(I) = \gcd\{\mathrm{nrd}(x) : x \in I\}.$$

Deuring proved that there is almost an equivalence of categories between supersingular elliptic curves over $\mathbb{F}_{p^2}$ (and their isogenies) and maximal orders in the quaternion algebra $B_{p,\infty}$, in which $i^2 = -1$ and $j^2 = -p$ (and left-ideals connecting these orders).

We just need one more definition. Let $E/\mathbb{F}_{p^2}$ be a supersingular elliptic curve. Then $\pi : (x, y) \mapsto (x^p, y^p)$ is the *Frobenius isogeny* from E. When E is defined over $\mathbb{F}_p$ we have seen this already as the Frobenius enodmorphism, but when E is defined over $\mathbb{F}_{p^2}$ (but not $\mathbb{F}_p$) it defines an isogeny. The codomain of π is called the *Frobenius conjugate* of E and is denoted by $E^{(p)}$.

Now, Deuring showed that there is a 1-1 correspondence between pairs $(E, E^{(p)})$ of Frobenius-conjugate supersingular elliptic curves and maximal orders in $B_{p,\infty}$ via the map $E \mapsto \mathrm{End}(E)$. Furthermore, under this correspondence, ℓ-isogenies map to left-ideals of norm ℓ, where kernels are associated with ideals as we saw in the context of CSIDH.

This means that we can translate any problem on a supersingular isogeny graph to a 'quaternion order graph', where the vertices are maximal orders in a quaternion algebra and the edges are left-ideals.

Our isogeny-based cryptosystems have all been based on the hardness of 'the isogeny problem', which can be stated as: given uniformly random supersingular elliptic curves E and $E'/\mathbb{F}_{p^2}$, compute an isogeny between them. In the quaternion graph, this problem becomes: given uniformly random maximal orders $\mathcal{O}$ and $\mathcal{O}'$ in $B_{p,\infty}$, compute a left-ideal I of $\mathcal{O}$ such that $\mathcal{O}' = \mathcal{O}_r(I)$. It turns out that this problem can be solved in polynomial-time using the KLPT algorithm due to Kohel, Lauter, Petit, and Tignol: and is how the verification arrow in SQISign is computed. Alice starts from an E_0 with known endomorphism ring, and computes the endomorphism rings of all the elliptic curves in the diagram via the Deuring correspondence and the knowledge of the isogenies. She can then run the KLPT algorithm to find an ideal connecting $\mathrm{End}(E_{\mathrm{pk}})$ and $\mathrm{End}(E_{\mathrm{ver}})$, and translate it back to an isogeny via the Deuring correspondence.

It remains to mention that for all this to lead to a secure system it is fundamental that the attacker cannot also compute these endomorphism rings, and in fact Wesolowski has proved that the supersingular isogeny problem is polynomially equivalent to the hard problem of computing the endomorphism ring of a uniformly random supersingular elliptic curve.