2 Elliptic curves as abelian groups

In Lecture 1 we defined an elliptic curve as a smooth projective curve of genus 1 with a distinguished rational point. An equivalent definition is that an elliptic curve is an abelian variety of dimension one. An *abelian variety* is a smooth projective variety that is also a group, where the group operation is defined by rational functions (ratios of polynomials). Remarkably, these conditions (in particular, the fact that we are working with projective varieties) force the group to be commutative, which is why they are called abelian varieties.

A variety is (roughly speaking) the zero locus of a set of polynomials, subject to an irreducibility condition. The precise definition won't concern us here, it is enough to know that a variety of dimension one is a curve, so an abelian variety of dimension one is a smooth projective curve with a group structure specified by rational functions. We will prove in this lecture that elliptic curves are abelian varieties. In fact the converse holds, every abelian variety of dimension one is an elliptic curve, but we won't prove this.

As mentioned in the first lecture, it is possible to associate an abelian variety to any smooth projective curve; this abelian variety is called the *Jacobian* of the curve. The dimension of the Jacobian is equal to the genus of the curve, which means that Jacobian is typically a much more complicated object than the curve itself, which has dimension one. Writing explicit equations for the Jacobian as a projective variety is quite complicated in general, but for elliptic curves, the curve and its Jacobian both have dimension one, and in fact they are isomorphic as projective varieties.

2.1 The group law for Weierstrass curves

Recall from Lecture 1 that the group law for an elliptic curve defined by a Weierstrass equation is determined by the following rule:

Three points on a line sum to zero, which is the point at infinity.

For convenience we will assume we are working over a field k whose characteristic is not 2 or 3, so that we may assume we are working with an elliptic curve E/k defined by a Weierstrass equation of the form

$$E : y^2 = x^3 + Ax + B.$$

The case of a general Weierstrass equation $y^2+a_1xy+a_3y = x^3+a_2x^2+a_4x+a_6$ is essentially the same, but the formulas are slightly more complicated; see [4, III.2.3] for details.

Recall that although we typically may our curves using an affine equation in the variables x and y, we are really working with the corresponding projective curve, which in this case is given by the homogeneous equation

$$E : y^2z = x^3 + Axz^2 + Bz^3$$

In order to specify an elliptic curve we need not only an equation defining the curve, but also a distinguished rational point, which acts as the additive identity 0. For curves in Weierstrass form choose the point $O := (0 : 1 : 0)$, which is the unique point on the curve E that lies on the line $z = 0$ at infinity: if $z = 0$ then $x = 0$ and we may assume $y = 1$ after scaling the projective point $(0 : y : 0)$ by $1/y$ (note that $x = z = 0$ forces $y \neq 0$; by definition, $(0 : 0 : 0)$ is not a projective point)

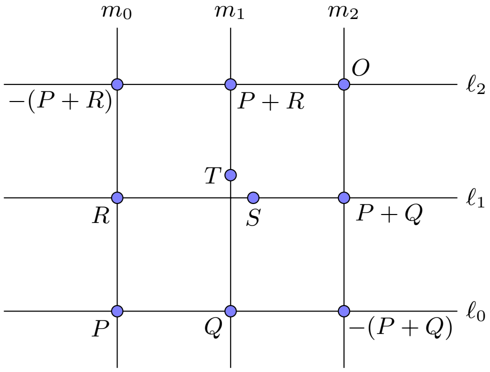
Every point $P \neq 0$ on the curve E thus has a nonzero z-coordinate which we can scale to be 1, and we use $P = (x_0, y_0) := (x_0 : y_0 : 1)$ to such an affine point. Notice that the point $Q = (x_0, -y_0)$ also lies on the curve E, and the projective line through P and Q is defined by $x = x_0 z$, which also passes through $O = (0 : 1 : 0)$. The three points P, Q, O lie on a line, so by the definition of the group law, $P + Q + O = P + Q = O$; thus $Q = -P$. We can also check that O acts as the identity element: the line between O and any point P intersects the curve at $-P$ (this is a double intersection at a tangent when $P = -P$). We then have $O + P + (-P) = O$, so $O + P = P$. Commutativity of the group law follows immediately from our definition, so $P + O = P$ also holds.

Associativity is not obvious, and while it can be rigorously proven algebraically, this is a tedious task that does not yield much insight. So we will give two proofs. The first will only apply to the generic case but it is short and provides some explanation as to *why* the group operation is associative. The second will be algebraic and fully rigorous, but we will let Sage do all the dirty work for us.

2.1.1 A geometric proof of associativity in the generic case

This is an adaptation of the proof in [2, p. 28]. Let P, Q, R be three points on an elliptic curve E over a field k that we may assume is algebraically closed (if the group law is associative over $\bar{k}$ then it is certainly also associative when we restrict to k). We shall also assume that P, Q, R, and the zero point O are all in *general position* (this means that in the diagram below there are no relationships among the points other than those that necessarily exist by construction).

The line ℓ_0 through P and Q meets the curve E at a third point, $-(P+Q)$, and the line m_2 through O and $-(P+Q)$ meets E at $P+Q$. Similarly, the line m_0 through P and R meets E at $-(P+R)$, and the line ℓ_2 through O and $-(P+R)$ meets E at $P+R$. Let S be the third point where the line ℓ_1 through $Q+P$ and R meets E, and let T be the third point where the line m_1 through Q and $P+R$ meets E. See the diagram below.



We have $S = -((Q+P)+R)$ and $T = -(Q+(P+R))$. It suffices to show $S = T$. Suppose not. Let $g(x, y, z)$ be the cubic polynomial formed by the product of the lines ℓ_0, ℓ_1, ℓ_2 in homogeneous coordinates, and similarly let $h(x, y, z) = m_0 m_1 m_2$. We may assume $g(T) \neq 0$ and $h(S) \neq 0$, since the points are in general position and $S \neq T$. Thus g and h are linearly independent elements of the k-vector space V of homogeneous cubic polynomials in $k[x, y, z]$. The space V has dimension $\binom{3+2}{2} = 10$, thus the subspace of homogeneous cubic polynomials that vanish at the eight points O, P, Q, R, $\pm(Q+P)$, and $\pm(P+R)$ has dimension 2 and is spanned by g and h. The polynomial $f(x, y, z) = x^3 + Axz^2 + Bz^3 - zy^2$ that defines E is a nonzero element of this subspace, so we may write $f = ag + bh$ as a linear combination of g and h. Now $f(S) = f(T) = 0$, since S and T are both points on E, but $g(S) = h(T) = 0$ and $g(T), h(S) \neq 0$, which implies that both a and b are zero. But this is a contradiction because f is not the zero polynomial.

This completes our geometric proof of the group law (in the generic case). In order to give a completely general algebraic proof, and to be able to actually perform group operations explicitly, we need explicit formulas for computing the sum of two points.

2.1.2 The group law in algebraic terms

Let P and Q be two points on our elliptic curve $E\colon y^2 = x^3 + Ax + B$. We want to compute the point $R = P + Q$ by expressing the coordinates of R as rational functions of the coordinates of P and Q. If either P or Q is the point O at infinity, then R is simply the other point, so we assume that P and Q are affine points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$. There are two cases.

Case 1. $x_1 \neq x_2$. The line $\overline{PQ}$ has slope $m = (y_2 - y_1)/(x_2 - x_1)$, which yields the linear equation $y - y_1 = m(x - x_1)$ for $\overline{PQ}$. This line is not vertical, so it intersects the curve E in a third affine point $-R = (x_3, -y_3)$. Plugging the equation for the line $\overline{PQ}$ into the equation for the curve E yields

$$(m(x - x_1) + y_1)^2 = x^3 + Ax + B.$$

Expanding the LHS and moving every term to the RHS yields a cubic equation

$$g(x) := x^3 - m^2x^2 + \cdots = 0,$$

where the ellipsis hides lower order terms in x. The monic cubic polynomial $g(x)$ has two roots $x_1, x_2 \in k$ and therefore factors in $k[x]$ as

$$g(x) = (x - x_1)(x - x_2)(x - x_3),$$

where $x_3 \in k$ is the x-coordinate of the third point $-R$ on the intersection of $\overline{PQ}$ and E. Comparing the coefficient of x^2 in the two expressions above for $g(x)$ shows that $x_1+x_2+x_3 = -m^2$, and therefore $x_3 = m^2 - x_1 - x_2$. We can then compute the y-coordinate $-y_3$ of $-R$ by plugging this expression for x_3 into the equation for $\overline{PQ}$, and we have

$$\begin{aligned} m &= (y_2 - y_1)/(x_2 - x_1), \\ x_3 &= m^2 - x_1 - x_2, \\ y_3 &= m(x_1 - x_3) - y_1, \end{aligned}$$

which expresses the coordinates of $R = P + Q$ as rational functions of the coordinates of P and Q as desired. To compute $P + Q = R$, we need to perform three multiplications (one of which is squaring m) and one inversion in the field k. We'll denote this cost $3\mathbf{M}+\mathbf{I}$; we are ignoring the cost of additions and subtractions because these are typically negligible compared to the cost of multiplications and (especially) inversions.

Case 2. $x_1 = x_2$. We must have $y_1 = \pm y_2$. If $y_1 = -y_2$ then $Q = -P$ and $P + Q = R = 0$. Otherwise $P = Q$ and $R = 2P$, and the line $\overline{PQ}$ is the tangent to P on the equation for E, whose slope we can compute by implicit differentiation. This yields

$$2y\,dy = 3x^2dx + A\,dx,$$

so at the point $P = (x_1, y_1)$ the slope of the tangent line is

$$m = \frac{dy}{dx} = \frac{3x_1^2 + A}{2y_1},$$

and once we know m we can compute x_3 and y_3 as above. Note that we require an extra multiplication (a squaring) to compute m, so computing $R = 2P$ has a cost of $4\mathbf{M}+\mathbf{I}$.

Remark 2.1. You might object that we have not formally defined implicit differentiation over an arbitrary field, nor have we shown that this gives us the slope of the tangent line. One can rigorously justify this (using Kahler differentials, for example), but it is easy to verify that it works in our case: if you plug $y = m(x - x_1) + y_1$ into the curve equation $E\colon y^2 = x^3 + Ax + B$ using the slope $m = (2x_1^2 + A)/2y_1$ we computed using implicit differentiation, you will find that x_1 is a double root, and since the point $(x_1, -y_1)$ does not lie on the line $L\colon y = m(x - x_1) + y_1$ unless $y_1 = 0$, the point (x_1, y_1) has multiplicity 2 in the intersection $E \cap L$, which implies that L is tangent to E at (x_1, y_1) as claimed.

With these equations in hand, we can now prove associativity as a formal identity, treating $x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3, A, B$ as indeterminants subject to the three relations implied by the fact that P, Q, R lie on the curve E. See the Sage worksheet

Lecture 2 Proof of associativity

for details, which includes checking all the special cases.

The equations above can be converted to projective coordinates by replacing x_1, y_1, x_2, and y_2 with x_1/z_1, y_1/z_1, x_2/z_2, and y_2/z_2 respectively, and then writing the resulting expressions for x_3/z_3 and y_3/z_3 with a common denominator. When $P \neq Q$ we obtain

$$\begin{aligned}
x_3 &= (x_2z_1 - x_1z_2)((y_2z_1 - y_1z_2)^2 z_1z_2 - (x_2z_1 - x_1z_2)^2(x_2z_1 + x_1z_2))\\
y_3 &= (y_2z_1 - y_1z_2)((x_2z_1 - x_1z_2)^2(x_2z_1 + 2x_1z_2) - (y_2z_1 - y_1z_2)^2 z_1z_2) - (x_2z_1 - x_1z_2)^3 y_1z_2\\
z_3 &= (x_2z_1 - x_1z_2)^3 z_1z_2
\end{aligned}$$

and for $P = Q$ we obtain

$$\begin{aligned}
x_3 &= 2y_1z_1(A^2(z_1^2 + 3x_1^2)^2 - 8x_1y_1^2z_1)\\
y_3 &= A(z_1^2 + 3x_1^2)(12x_1y_1^2z_1 - A^2(z_1^2 + 3x_1^2)^2) - 8y_1^4z_1^2\\
z_3 &= (2y_1z_1)^3
\end{aligned}$$

These formulas are more complicated, but they have the advantage of avoiding inversions, which are more costly than multiplications (in a finite field of cryptographic size inversions may be 50 or even 100 times more expensive than multiplications). With careful reuse of common subexpressions these formulas lead to a cost of $12\mathbf{M}$ for addition (of distinct points) and $14\mathbf{M}$ for doubling.

2.2 Edwards curves

Many alternative representations of elliptic curves have been proposed over the years that lead to different formulas for the group law. We give just one example here, Edwards curves [1, 3], which have two significant advantages over Weierstrass equations. Let d be a non-square element of a field k whose characteristic is not 2. Then the equation

$$x^2 + y^2 = 1 + dx^2y^2 \qquad (1)$$

defines an elliptic curve with distinguished point $(0, 1)$.

Remark 2.2. The plane projective curve defined by equation (1) has two singular points at infinity, violating our requirement that an elliptic curve be smooth. However, this plane curve can be *desingularized* by embedding it in $\mathbb{P}^3(k)$. The points at infinity are then no longer rational, and do not play a role in the group operation on $E(k)$, whose elements can all be uniquely represented as solutions (x, y) to equation (1) above.

The group operation is given by

$$(x_3, y_3) = \left(\frac{x_1y_2 + x_2y_1}{1 + dx_1x_2y_1y_2}, \frac{y_1y_2 - x_1x_2}{1 - dx_1x_2y_1y_2}\right), \tag{2}$$

which implies that the inverse of (x_1, y_1) is $(-x_1, y_1)$. In contrast to the formulas for curves in Weierstrass form, the formula in (2) is well defined for every pair of points (x_1, y_1) and (x_2, y_2) in $E(k)$.

To prove this, let us suppose for the sake of obtaining a contradiction that one the denominators in (2) is zero (making the formulas undefined). Then we must have

$$(1 + dx_1x_2y_1y_2)(1 - dx_1x_2y_1y_2) = 1 - d^2x_1^2x_2^2y_1^2y_2^2 = 0,$$

so $d^2x_1^2x_2^2y_1^2y_2^2 = 1$, and therefore x_1, x_2, y_1, y_2 are all nonzero. Applying this and the curve equation (twice) yields

$$x_1^2 + y_1^2 = 1 + dx_1^2y_1^2 = 1 + \frac{1}{dx_2^2y_2^2} = \frac{x_2^2 + y_2^2}{dx_2^2y_2^2}.$$

By adding or subtracting $2x_1y_1 = \pm 2/(dx_2y_2)$ to both sides we can obtain

$$(x_1 \pm y_1)^2 = \frac{(x_2 \pm y_2)^2}{dx_2^2y_2^2},$$

with either choice of sign on the LHS (the sign on the RHS may vary, but in any case the numerator of the RHS is a square). Since x_1 and y_1 are nonzero, one of $x_1 + y_1$ and $x_1 - y_1$ is nonzero, and this implies that d is a square in k, but this is a contradiction, since we assumed from the beginning that c is not a square.

As written, the group law involves five multiplications and two inversions (ignoring the multiplication by d, which we can choose to be small), which is greater than the cost of the group operation in Weierstrass form. However, in projective coordinates we have

$$\frac{x_3}{z_3} = \frac{z_1z_2(x_1y_2 + x_2y_1)}{z_1^2z_2^2 + dx_1x_2y_1y_2}, \qquad \frac{y_3}{z_3} = \frac{z_1z_2(y_1y_2 - x_1x_2)}{z_1^2z_2^2 - dx_1x_2y_1y_2}.$$

There are a bunch of common subexpressions here, and in order to compute z_3, we need a common denominator. Let $r = z_1z_2$, let $s = x_1y_2 + x_2y_1$, let $t = dx_1y_2x_2y_1$, and let $u = y_1y_2 - x_1x_2$. We then have

$$x_3 = rs(r^2 - t), \qquad y_3 = ru(r^2 + t), \qquad z_3 = (r^2 + t)(r^2 - t).$$

This yields a cost of 12**M**. If we compute s as $s = (x_1 + y_1)(x_2 + y_2) - x_1x_2 - y_1y_2$, the cost is reduced to 11**M**.

A simple Sage implementation of these formulas can be found here:

Lecture 2 Group law on Edwards curves

Because the expression in (2) is well-defined at every point in $E(k)$, we do not need separate formulas for addition and doubling. Moreover, we don't even need to check the cases where one or both points is the identity element, or one is the negation of the other, the same formula works in every case. Such formulas are said to be *complete*, and they have two distinct advantages. First, they can be implemented very efficiently as a straight-line program with no branching. Second, they protect against what is known as a *side-channel* attack. If you are using different formulas for addition and doubling, it is possible that an adversary may be able to externally distinguish these cases, e.g. by monitoring the CPU (electronically, thermally, or even acoustically) and noticing the difference in the time required or energy used by each operation. They can then use this information to break a cryptosystem that performs scalar multiplication nP by an integer n that is meant to be secret (as in Diffie-Hellman key exchange, for example), because they can the sequence of doubling-add-adding used in scalar multiplication effectively encodes the binary representation of n. Using a complete formulas prevents a side channel attack because exactly the same sequence of instruction is executed for every group operation.

Having said that, if you know you want to double a point and are not concerned about a side-channel attack, there are several optimizations that can be made to the formulas above (these include replacing $1 + dx^2y^2$ with $x^2 + y^2$). This reduces the cost of doubling on an Edwards curves to 7**M**, half the 14**M** cost of doubling a point in Weierstrass coordinates.

The explicit formulas database contains optimized formulas for Edwards curves and various generalizations, as well as many other forms of elliptic curves. Operation counts and verification scripts are provided with each set of formulas.

We should note that, unlike Weierstrass equations, not every elliptic curve can be defined by an equation in Edwards form. In particular, an Edwards curve always has a rational point of order 4, the point $(1, 0)$, but most elliptic curves do not have a rational point of order 4.

References

[1] Daniel J. Bernstein and Tanja Lange, *Faster addition and doubling on elliptic curves*, Advances in Cryptology - ASIACRYPT 2007, Lecture Notes in Computer Science **4833**, Springer-Verlag, New York (2007), 29–50.

[2] J. W. S. Cassels, *Lectures on elliptic curves*, London Mathematical Society Student Texts **24**, Cambridge Universtity Press, 1991.

[3] Harold M. Edwards, *A normal form for elliptic curves*, Bulletin of the American Mathematical Society **44** (2007), 393–422.

[4] Joseph H. Silverman, *The arithmetic of elliptic curves*, 2nd ed., Springer, 2009.