Elliptic Curve Cryptography

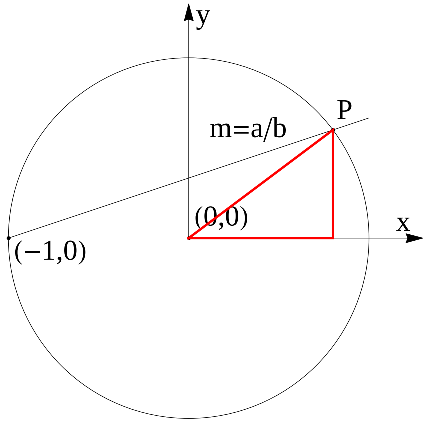
November 3, 2013

1 A Warmup Problem

We'll begin by looking at a problem whose solution will illustrate some of the techniques used in elliptic curve cryptography, but which involves algebra that is much simpler. Our goal will be to find a general formula for pythagorean triples; that is, sets of integers $\{x, y, z\}$ that satisfy the formula: $x^2 + y^2 = z^2$. Another way of stating the problem is that we are looking for sets of three integers that form the sides of a right triangle, where x and y are the lengths of the legs and z is the length of the hypotenuse. You're probably familiar with some of them, like $\{3, 4, 5\}$, $\{5, 12, 13\}$ and $\{8, 15, 17\}$.

We're going to employ a trick; namely, that for a quadratic equation, there are usually two roots, and those roots correspond to the points where a straight line passes through a conic section (in this case, a circle). The trick we will use is that we will use a line that passes through a particular known point, and that will make the factorization of the quadratic equation easy, since we know one of the factors. Once it's factored, we obtain the other root, which will yield the information we want. The other interesting part is that after we've factored out the known root, the equation will be linear with rational coefficients, so the second root will also have rational coefficients.

Exercise: Prove that the three sets of integers in the previous paragraphs satisfy the condition of being pythagorean triples.



In the figure above, we have the unit circle ($x^2 + y^2 = 1$) and a line passing through the point $(-1, 0)$ and having a rational slope of $m = a/b$. This line will intersect the circle at a point P. We will show that this generates right triangles with rational coordinates with hypotenuse OP with one leg being on the x-axis. Since the side lengths are rational, they can be multiplied by a constant to make a triangle with integer sides.

Let's just look for triangles in the first quadrant, meaning that the slope of the line passing through $(-1, 0)$ and P must have a slope between 0 and 1, meaning that $a < b$.

The equation for the line is $y = (a/b)(x - 1)$, so to find the coordinates of P we need only substitute that value of y into the equation for the unit circle to find the intersection:

$$x^2 + \Big(\frac{a}{b}(x+1)\Big)^2 = 1.$$

This is a quadratic equation, but we know one of the roots: $x = -1, y = 0$ since that's one of the two places the line intersects the circle. Here is how to solve for x, which we can then substitute into the equation for the line to obtain y:

$$\begin{aligned}
x^2 + \Big(\frac{a}{b}(x+1)\Big)^2 &= 1 \\
(x^2 - 1) + \Big(\frac{a}{b}(x+1)\Big)^2 &= 0 \\
(x+1)(x-1) + \Big(\frac{a^2}{b^2}\Big)(x+1)^2 &= 0 \\
(x-1) + \Big(\frac{a^2}{b^2}\Big)(x+1) &= 0 \\
b^2(x-1) + a^2(x+1) &= 0 \\
(b^2 + a^2)x &= (b^2 - a^2) \\
x &= \frac{b^2 - a^2}{b^2 + a^2} \\
y &= \Big(\frac{a}{b}\Big)(x+1) = \Big(\frac{a}{b}\Big)\Big(\frac{2b^2}{b^2 + a^2}\Big) = \frac{2ab}{b^2 + a^2}
\end{aligned}$$

Since $x^2 + y^2 = 1$, it's easy to check our work.

Exercise: Verify that:

$$\Big(\frac{b^2 - a^2}{b^2 + a^2}\Big)^2 + \Big(\frac{2ab}{b^2 + a^2}\Big)^2 = 1.$$

Now we can just multiply by $b^2 + a^2$ and obtain integer values for the lengths of the sides:

$$\{b^2 - a^2, 2ab, b^2 + a^2\}.$$

Since any rational point on the circle in the first quadrant will form a line through $(-1, 0)$ with a rational slope, the equation above will yield all possible pythagorean triples.

Exercise: Try substituting various values of a and b into the formula above to obtain a few examples of pythagorean triples. Find a and b that yield the "standard" triangles, $\{3, 4, 5\}$, $\{5, 12, 13\}$ and $\{8, 15, 17\}$. Find a few more, just for fun.

2 Elliptic curves

We will now do something similar, but with elliptic curves. (The word “elliptic” will not mean that the equations are ellipses.) The general form of the elliptic curve equation is this:

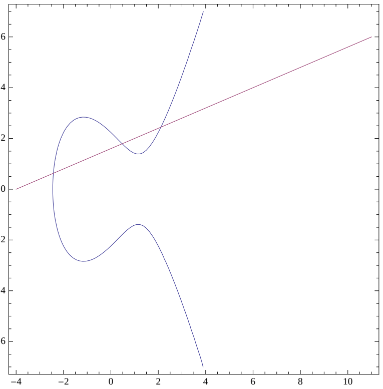
$$ay^2 + by = cx^3 + dx^2 + ex + f,$$

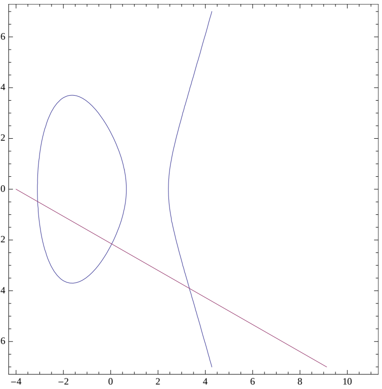
where a, b, c, d, e and f are real numbers.

Using linear substitutions for x and y we can convert this equation to one of the form:

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

Exercise: Try to find substitutions for x and y that will do this. The substitutions will look something like $x' = \alpha x + \beta$ and $y' = \gamma y + \delta$. Hint: You can do it in multiple steps: First divide through by c to eliminate the coefficient on x^3, then if the new coefficient of y^2 is g, substitute $y = \sqrt{g}y'$, et cetera.





The figures above (ignore the straight lines for now) show the curves represented by the equations $y^2 = x^3 - 4x + 5 = 0$ (on the left) and $y^2 = x^3 - 8x + 5 = 0$ (on the right).

With quadratic equations, the curve is intersected by any line in zero, one or two places. When it is intersected in one place, the line is touching (tangent to) the curve. With a general cubic equation, there are up to three intersections, and in the case of the equations above, when there are one or two intersections, there is a tangency of the line to the curve. The straight lines in the illustrations above show how you can obtain three intersections.

Note that when a line is tangent to a curve, it is effectively a double root, meaning that the equation can be factored with two copies of the same term. Note also that if (x, y) lies on the curve, then so also does $(x, -y)$, since the only way that y enters the equation is in the form y^2.

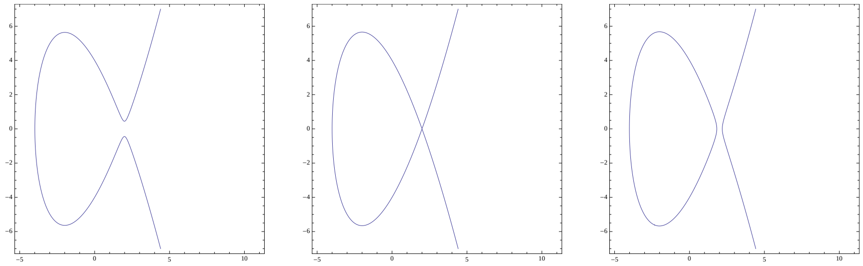
If we can find two points on such a curve that have rational coefficients, then the straight line passing through those two will pass through the curve at a third point and we can factor out the two roots so that the remaining equation will effectively be linear, and will thus also have rational coordinates.

3 Singular Elliptic Curves

For the general elliptic curve $y^2 = x^3 + Ax + B$ to behave “nicely” it is important that the curve not be singular. The condition to avoid a singularity is simply this:

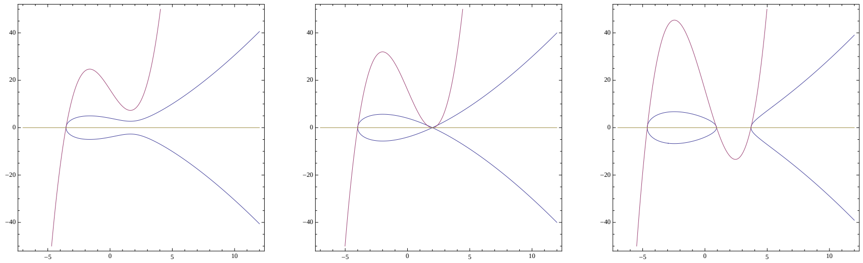
$$4A^3 + 27B^2 \neq 0.$$

The following three plots illustrate what a singularity looks like. One particular set of values that will produce a singularity is if $A = -12$ and $B = 16$. The following three plots, from left to right, differ only in the value of A; B is always 16. On the left, $A = -11.9$, in the center, $A = -12$ and on the right, $A = -12.1$



Here’s what is going on: In the graphs below, the purple line is the graph of $y = x^3 + Ax + B$ and the blue line is the graph of $y^2 = x^3 + Ax + B$. The value of $x^3 + Ax + B$ must be positive before we can take a square root, so if the purple line is negative, there is no blue curve, and when there is a blue curve, it will include both the positive and negative values of the square root.

If the purple line is tangent to the x-axis, that’s when we get a singularity.



There will be a point of tangency when the minimum of the curve $y = x^3 + Ax + B$ is on the x-axis. The minimum will occur when the derivative, $3x^2 + A = 0$, or when

$x = \sqrt{-A/3}$. For this to be touching the x-axis, we need:

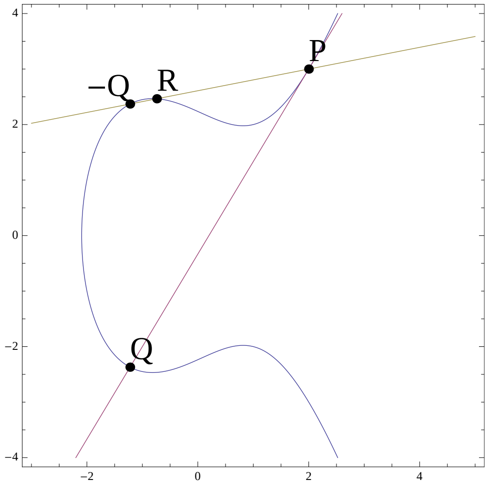
$$\begin{aligned}
\Big(\sqrt{-\frac{A}{3}}\Big)^3 + \sqrt{-\frac{A}{3}}A + B &= 0 \\
\sqrt{-\frac{A}{3}}\Big(-\frac{A}{3} + A\Big) &= -B \\
\sqrt{-\frac{A}{3}}\Big(\frac{2A}{3}\Big) &= -B \\
-\frac{4A^3}{27} &= B^2 \\
0 &= 4A^3 + 27B^2
\end{aligned}$$

For this reason, from now on, we'll assume that $4A^3 + 27B^2 \neq 0$.

4 Elliptic Curve Example

In this example, we'll show how a single point on a curve can be used to generate other points on the curve.

Consider the curve with $A = -2$ and $B = 5$: $y^2 = x^3 - 2x + 5$. We see that the point $P = (2, 3)$ lies on the curve.



See the figure above. If we begin with P and draw the line that is tangent to the curve at the point P, it will intersect the curve at another point Q. Because Q is on the curve, the point $-Q$ (which is the same as Q, but with the negative of the y-coordinate) is also on the curve. Now we can draw the line through $-Q$ and P to obtain another point R. From there, we can find $-R$ or intersect the line QR with the curve to obtain more and more points. This process can be extended forever, and since the original coordinates of the point P were rational, all the others will be, too.

In this example, here are the coordinates of the generated points:

$$P = (2, 3), \quad Q = (-11/9, -64/27),$$

$$-Q = (-11/9, 64/27), \quad R = (-622/841, 60111/24389)$$

The coordinates are rational, but the rational numbers get messier and messier as we go on.

Exercise: Check that all the points above lie on the curve, and check that the points $-Q, R$ and P all lie on the same line.

5 Finite Fields

A field is a mathematical structure that includes a set of objects and two binary operations on those objects, addition ($+$) and multiplication ($\cdot$). Those operations satisfy a bunch of familiar axioms for all a, b and c:

- (Associativity) $(a + b) + c = a + (b + c), \quad (a \cdot b) \cdot c = a \cdot (b \cdot c)$
- (Commutativity) $a + b = b + a, \quad a \cdot b = b \cdot a$
- (Distributivity) $a \cdot (b + c) = a \cdot b + a \cdot c$
- (Additive Identity) There exists 0 such that: $a + 0 = a$
- (Multiplicative Identity) There exists 1 such that: $a \cdot 1 = a$
- (Additive Inverse) There exists $-a$ such that $a + (-a) = 0$
- (Multiplicative Inverse) If $a \neq 0$ there exists a^{-1} such that $a \cdot a^{-1} = 1$

We often omit the multiplication symbol and write $a \cdot b$ as ab. Similarly, we write things like $a + (-b)$ as $a - b$. We also often write a^{-1} as $1/a$.

Three fields with which you are almost certainly familiar are the real numbers, the complex numbers and the rational numbers, and all three of these fields have an infinite number of elements.

What may be surprising, however, is that there also exist finite fields.

Exercise Show that there is a field of size 2 containing just two elements, 0 and 1. The operations are all as you expect, except that $1 + 1 = 0$ in this field. Check that all the field properties listed above are satisfied.

Without proof, we state the following fact: If p is any prime number and n is any positive integer, then there exists a finite field of size p^n. There are no other finite fields.

In this article, we will only consider finite fields of size p, where p is a prime number. In such a field, addition and multiplication are just ordinary addition and multiplication, but taken modulo p.

Consider the example where $p = 7$. Listed below are the addition and multiplication tables for the finite field of order 7. Things like associativity, commutativity and distributivity follow from the fact that addition and multiplication of the integers satisfy those properties. The fact that there's an additive and multiplicative inverse for appropriate numbers can be verified by making sure that (for addition) the number 0 appears in each row and column. To verify the inverse we simply check that, ignoring the row and column of zeroes, that the number 1 appears in every row and column.

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

×	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

Many many features of the fields you're familiar with (reals, rationals or complex numbers) carry over into finite fields. We can, for example, form polynomials, do factorization, et cetera, and so it makes perfect sense to consider equations of the form:

$$y^2 = x^3 + Ax + B$$

in the finite field, where x, y, A and B are taken to be numbers in the finite field and the addition, multiplication (and exponentiation, which is just repeated multiplication) are performed using the tables for that finite field.

In the finite field with 7 elements, let's consider the "curve" where $A = 1$ and $B = 6$: $y^2 = x^3 + x + 6$. The word "curve" is in quotes since there are only a finite set of points (x, y), so a plot would not be a curve, but rather a set of points. In fact, with 7 elements in the field, there are 49 possible points (x, y). It turns out that all of the following ten points lie on the curve (satisfy the equation):

(1, 1)	(1, 6)	(2, 3)	(2, 4)	(3, 1)
(3, 6)	(4, 2)	(4, 5)	(6, 2)	(6, 5)

Exercise: Check to see that some of the points above do satisfy the equation (lie on the "curve"). Note that we pointed out earlier that if (x, y) satisfies the equation, then so does $(x, -y)$. We see this occurring, for example, in that $(2, 3)$ and $(2, 4)$ work. This is because $-4 = 3$ (modulo 7), since $4 + 3 = 7$, et cetera.

Exercise: Show that the points $(3, 1)$, $(2, 3)$ and $(6, 2)$ lie on the same line. Remember that the equation of a line in a finite field will look exactly like the equation you're used to; something like: $Ax + By = C$, where A, B and C are numbers in the finite field. (You can use formulas like the point-point or point-slope forms for the equation of a line. Just remember that when you need to do division, the division is done using the inverse of the multiplication tables for the field.) So if the slope turned out to be, say,

$5/3$, you'd need to find the inverse of 3, which is the number, which when multiplied by 3, yields 1. That would be 5 in this case. Since $1/3 = 5$, then $5/3 = 5 \cdot 5 = 4$.

If you try picking pairs of points at random from the list above and try to find the third point on the line, this will work sometimes, but it can fail for a couple of reasons. The first is that the line you've picked is effectively "tangent" to the "curve". What this really means, algebraically, is that there is a double root when you try factoring the equation. The other thing that can happen is that you pick two points that have the same x-coordinate (and therefore have y-coordinates are are negatives of each other). In the continuous case of real curves, this is like selecting a vertical line which has no more points on it. What we do, mathematically, is imagine that the unbounded ends of the curve actually "go to infinity" and we add an artificial "point at infinity" that we'll call $\mathcal{O}$. We will say that this is the third point on any vertical line (line passing through two points with the same x-coordinate).

With this artificial point $\mathcal{O}$ we now have a nice set of points that satisfy the equation. Any pair of them determine a third, although that third may be the point $\mathcal{O}$, or it may be one of the original two points, where that point will be a double root of the equation.

6 Elliptic Curve Mathematics

We'll work with this curve:

$$y^2 = x^3 + Ax + B$$

but unless we say so, we won't assume that we are working over any particular field, so our results will be as general as possible[1].

We assume that the curve is non-singular:

$$4A^2 + 27B^3 \neq 0.$$

We will also need to introduce a special point "at infinity" which we will call $\mathcal{O}$. It's easy to justify the fact that this point satisfies the equation. You can think of it as having coordinates (∞, ∞) since if you plug the two infinities into the curve you get $\infty = \infty$, which is "reasonable".

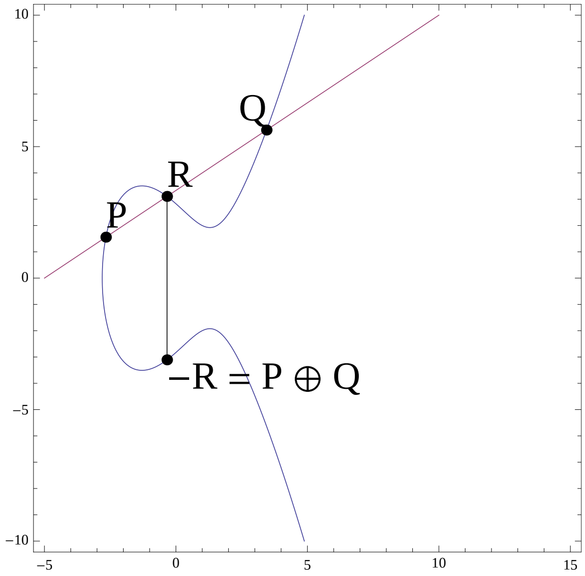
Let $\mathbb{E}$ be the set of all points that satisfy the equation of the elliptic curve. We will include $\mathcal{O}$ in this set. Thus in the finite field example in Section 5 the set $\mathbb{E}$ would consist of 11 points: the 10 finite points and the point at infinity. The other 39 points in the "plane" of the finite field with seven elements do not satisfy the equation we considered, so they are not of interest. Of course if we change A and/or B we would have a different equation and the set $\mathbb{E}$ would also change.

What we'd like to do is to introduce a notion of "addition" to the points that satisfy the elliptic curve equation, based on the geometry. Think of the geometry of the curves we've drawn up to now in the real coordinate plane and refer back to some of those illustrations for motivation.

[1] Already we've made a small assumption: the general cubic function cannot be reduced to this form if the field is of characteristic 2 or 3. But once we have an equation like this, we're fine.

The idea is this: any two distinct points P and Q on the curve determine a straight line and we need to find a third point on that line. If the straight line is vertical (in other words, if the x-coordinates are the same) then the third point is $\mathcal{O}$. If the straight line is tangent to the curve at one of the two points (algebraically, if the polynomial factors to yield a double root) then the third point is the point having the double root. The usual situation is that the line passes through the curve at the required third point (and it's impossible for a line to pass through more than three points). If one of the two points is $\mathcal{O}$ and the other is P, then the third point is $-P$, where $-P$ is the same as P except with the opposite y-coordinate.

The sum of two points in $\mathbb{E}$ is defined as follows: If R is the third point (as defined above) on the line PQ we will say that $P \oplus Q = -R$. Note that this definition even makes sense if we add a point P to itself, since that will be a double point on the curve, and the third point will be where the tangent line at P intersects the curve. The following figure illustrates the operation of adding points P and Q to obtain $P \oplus Q$:



The following table lists the properties of addition in $\mathbb{E}$. For all P, Q and R in $\mathbb{E}$ we have:

$$\begin{aligned} P+Q &= Q+P \\ P+\mathcal{O} &= P, \quad \text{so } \mathcal{O}+\mathcal{O}=\mathcal{O} \\ P+(Q+R) &= (P+Q)+R \\ P+(-P) &= \mathcal{O}, \quad \text{so } -\mathcal{O}=\mathcal{O} \end{aligned}$$

An interesting feature of this addition is that it makes sense to define multiples of a point P as follows:

$$\begin{aligned}
0P &= \mathcal{O} \\
1P &= P \\
2P &= P \oplus P \\
3P &= P \oplus P \oplus P \\
4P &= P \oplus P \oplus P \oplus P \\
\ldots &= \ldots
\end{aligned}$$

Let's find formulas for $P \oplus Q$. Suppose $P = (x_1, y_1)$ and $Q = (x_2, y_2)$. Then if the line through P and Q is given by the equation $y = mx + b$, assuming that $P \neq Q$ we find that:

$$m = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{and} \quad b = y_1 - mx_1.$$

If $P = Q$ then we can just take the derivative at (x_1, y_1) to find the slope[2]:

$$\begin{aligned}
d(y^2) &= d(x^3 + Ax + B) \\
2ydy &= (3x^2 + A)dx \\
\frac{dy}{dx} &= \frac{3x^2 + a}{2y}
\end{aligned}$$

So if $P = Q$, at (x_1, y_1) we have $m = (3x_1^2 + A)/2y_1$ and $b = y_1 - mx_1$.

Now that we have the equation for the line, we can find the third point (x_3, y_3) on the line. We need to find the intersection of the curve $y^2 = x^3 + Ax + B$ with the line $y = mx + b$:

$$(mx + b)^2 = x^3 + Ax + B.$$

Since we know that (x_1, y_1), (x_2, y_2) and (x_3, y_3) are all solutions, we know that:

$$\begin{aligned}
0 &= x^3 + Ax + B - (mx + b)^2 \\
0 &= (x - x_1)(x - x_2)(x - x_3) \\
0 &= x^3 - (x_1 + x_2 + x_3)x^2 + (x_1x_2 + x_2x_3 + x_3x_1)x - x_1x_2x_3
\end{aligned}$$

By matching coefficients, we can conclude that $m^2 = (x_1 + x_2 + x_3)$ so we can conclude that $x_3 = m^2 - x_1 - x_2$ and therefore $y_3 = mx_3 + b$ and $P \oplus Q = (x_3, -y_3)$.

Here's a complete set of rules:

$$\begin{aligned}
\text{If } P \neq Q \text{ and } x_1 = x_2 : &\quad P \oplus Q = \mathcal{O} \\
\text{If } P = Q \text{ and } y_1 = y_2 = 0 : &\quad P \oplus Q = \mathcal{O} \\
\text{Otherwise} : &\quad P \oplus Q = (m^2 - x_1 - x_2, -m^3 + m(x_1 + x_2) - b)
\end{aligned}$$

[2] It turns out that this works perfectly for finite fields as well.

7 Finite Elliptic Curves

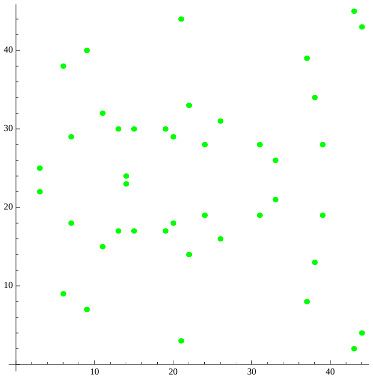
For their use in cryptography, we will want to use huge prime numbers p (perhaps hundreds of digits long) But let's try to get a better feel for the situation when p is of a moderate size; say, $p = 47$. All of the possible values in $\mathbb{E}$ will be chosen from the set $\{(a, b) : 0 \leq a, b \leq p - 1\}$. The size of $\mathbb{E}$ will depend on the values of A and B in the equation $y^2 = x^3 + Ax + B$.

7.1 A Small Example

If $p = 47$ and the equation is $y^2 = x^3 + 22x + 15$, then here is the set $\mathbb{E}$:

$$\begin{aligned}
\mathbb{E} \quad = \quad & \{\mathcal{O}, (3,22), (3,25), (6,9), (6,38), (7,18), (7,29), (9,7), \\
& (9,40), (11,15), (11,32), (13,17), (13,30), (14,23), (14,24), (15,17), \\
& (15,30), (19,17), (19,30), (20,18), (20,29), (21,3), (21,44), (22,14), \\
& (22,33), (24,19), (24,28), (26,16), (26,31), (31,19), (31,28), (33,21), \\
& (33,26), (37,8), (37,39), (38,13), (38,34), (39,19), (39,28), (43,2), \\
& (43,45), (44,4), (44,43)\}
\end{aligned}$$

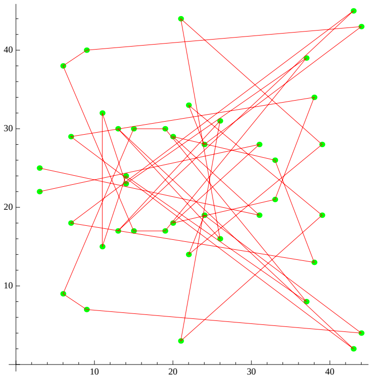
and the following is a plot of all those points, except, of course, for the point $\mathcal{O}$ which is "at infinity":



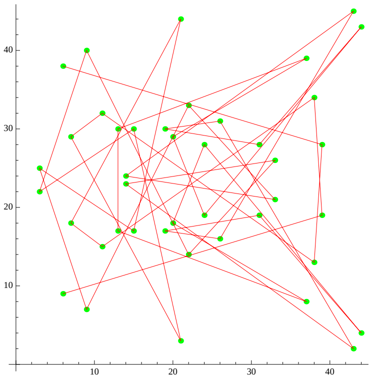
Notice that this plot is symmetric about the line $y = 23.5$ since if we work modulo 47 then $-y = 47 - y$. (There can sometimes be one non-symmetric point if a valid point has a y-coordinate of zero, since the symmetric point would be at 47, which is 0, modulo 47.)

Exercise: check to make sure that a few of the points listed in $\mathbb{E}$ above satisfy the equation $y^2 = x^3 + 22x + 15$.

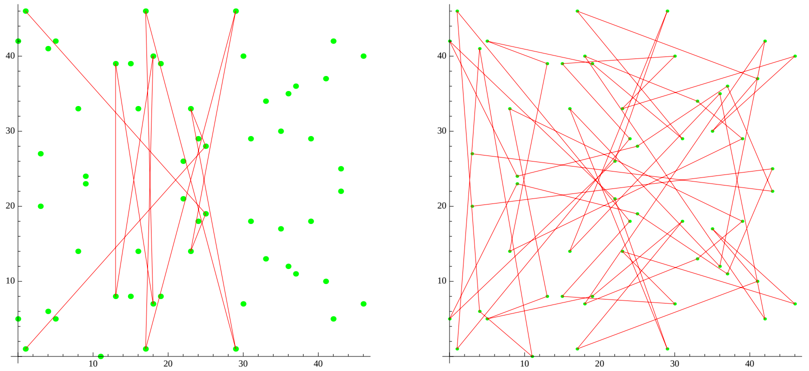
In the figure below, we begin with the point $P = (3, 22)$ and draw a line from P to $2P$ to $3P$, et cetera, and finish when we reach $-P = (3, 25)$. The next point will the $\mathcal{O}$ and then we would return to P, $2P$, and so on.



If we start from a different $P = (6, 9)$ and do the same thing, the path through the grid again hits all the points, but a completely-different looking path:



For the equation $y^2 = x^3 + 22x + 15$ every starting point P cycles through all the grid points, but this does not have to be the case. Again taking $P = 47$, let's look at the corresponding plots for the following equation: $y^2 = x^3 + 22x + 25$, starting at $P = (1, 1)$ (on the left) and starting at $P = (3, 27)$ (on the right). Notice that in this example, the point $(11, 0)$ satisfies the equation, so the plot is not completely symmetric here. If it were drawn on a cylinder (or torus) it would be symmetric since the points with coordinates 0 and 47 would lie on top of each other.



7.2 Calculation in a Finite Field

How were the equations for the lines calculated for these finite fields? The slope, m involves a division. One of the key features of being a field is that every element a except for zero has a multiplicative inverse: a number, which when multiplied by a, yields 1. Let's go through the calculation using the equation $y^2 = x^3 + 22x + 25$, and using the initial point $(x_1, y_1) = (1, 1)$ and the second point $(x_2, y_2) = (25, 28)$ we will show that the next point on the line is at $(x_3, y_3) = (23, 33)$, as is shown in the previous figure on the left.

From our previous work, we know that:

$$(x_3, y_3) = (m^2 - x_1 - x_2, -m^3 + m(x_1 + x_2) - b),$$

where $m = (y_2 - y_1)/(x_2 - x_1)$ and $b = y_1 - mx_1$. Since $m = (27/24)$ we need to find (effectively) $1/24$ modulo 47 which is 2, since $2 \cdot 24 = 48$ which is 1, modulo 47. Thus, (working modulo 47), we have $m = 54 = 7$ and $b = 1 - 7 \cdot 1 = -6 = 41$. So:

$$(x_3, y_3) = (49 - 1 - 25, -343 + 7 \cdot (26) - 41) = (23, 33).$$

7.3 Finding Inverses in a Finite Field

For the example above where $p = 47$ it is not too hard to find the inverse of a number. At worst, you simply test all 46 possibilities. But if p is huge, which it will be when we want to apply these elliptic curves to cryptography, checking through all the possibilities is simply not practical.

What comes to our aid, however, is Fermat's little theorem. If p is prime and $a \neq 0$ then:

$$a^{p-1} = 1(\bmod p).$$

Thus the inverse of a will simply be a^{p-2}. At first glance, this seems to be of no help at all, since if p is a 100-digit prime, raising a to a 100-digit power will take almost forever, and it will, if we simply multiply by a over and over again. But what if the exponent n of a^n were a perfect power of 2; if say, we want to calculate a^{128}? We can square a to obtain a^2, then square that to obtain a^4, square that to obtain a^8, then a^{16}, a^{32}, a^{64}, and finally, a^{128}. We can find a^{128} using just 7 multiplications instead of 127 of them. As the exponent gets larger and larger the savings gets huge. If the power were a 100-digit number, we can obtain it in fewer than 336 multiplications, which will take no time on a computer.

But what if the exponent is *not* a perfect power of 2? We're saved again, since any integer can be expressed in base-2. Here's the idea. Suppose we need to raise a to the power of 167. We write $167 = 128 + 32 + 4 + 2 + 1$. In the calculations from the previous paragraph, we know the values of a^1, a^2, a^4, a^{32} and a^{128}. Four more multiplications to combine those numbers will yield the desired result, since:

$$a^{167} = a^{128+32+4+2+1} = a^{128}a^{32}a^4a^2a^1.$$

For 100-digit exponents, we will need fewer than 671 multiplications: again, nothing on a modern computer.

8 A Trapdoor Problem

A trapdoor problem is one that is easy to do in one direction, but is very difficult to do in the opposite direction. An easy example is the problem of multiplying and factoring integers. If you have a decent computer, it is not a big deal at all to find a couple of 100-digit prime numbers and to multiply those together. But if someone gives you a 200-digit number and tells you that it is the product of two prime numbers, the job of finding those two seems to be very, very difficult.

This particular trap-door problem (multiplication/factoring) is used in another version of modern cryptography called RSA Encryption.

Discrete elliptic curves provide another type of trapdoor operation that can be used for encryption and the method is illustrated in Section 9.

With elliptic curves, the trapdoor problem is basically this: given the equation $y^2 = x^3 + Ax + B$, the prime number p and a point P that satisfies the equation, if p is huge and n is also a huge number (both with, say, 100 digits), then it is easy to calculate nP, but if you know only P and nP it seems to be very difficult to compute n.

Notice that we can use basically the same trick we used to find inverses in Section 7.3 to calculate the value of nP, even for very large values of n. From P, it's easy to obtain $2P$. Add two of those together to obtain $4P$. Adding two of those make $8P$, and continue to make $16P$, $32P$, $64P$ and so on. It only takes a few hundred steps (easy for a computer) to obtain powers of 2 in the coefficient that are hundreds of digits long.

If n is any number, it can be expressed as a sum of powers of 2, exactly as we did in Section 7.3. We can use exactly the same example: Suppose we wish to calculate $167P$. We find $1P$, $2P$, ..., $128P$ as described above, and then note that:

$$167P = (1 + 2 + 4 + 32 + 128)P = 1P + 2P + 4P + 32P + 128P,$$

so the entire calculation can be done with about a dozen point additions. For 100-digit values of n, we might need up to 600 or 700 point additions: again, nothing for a modern computer.

9 Elliptic Curve Ciphers

The main idea behind elliptic curve ciphers (ECCs) and behind many others is similar. First you convert your message to a sequence of 0's and 1's (perhaps simply by writing down the ASCII[3] text for it). Next, you and the person with whom you wish to communicate with somehow both obtain the same copy of a "key": a string of 0's and 1's that is at least as long as the message you wish to transmit and which is as random (unpredictable) as possible and you assure as well as possible that any opponents who want to read your message do not have a copy of this key.

Then the key is XORed, bit by bit, with your message and that encoded string is transmitted to the other person. That person simply applies the same XOR function to the encoded message and the original message is revealed.

We will look at every step of this procedure in detail in what follows.

9.1 Converting a Message to a Binary String

Assume that your message is written in standard English text, perhaps with digits, punctuation, et cetera. There is a standard way to encode English text to binary using the ASCII standard code (although any other one would work fine). Here is a table with the standard ASCII encoding that assigns a 7-bit pattern to every character:

	000	001	010	011	100	101	110	111
0000	ˆ@	ˆA	ˆB	ˆC	ˆD	ˆE	ˆF	ˆG
0001	ˆH	ˆI	ˆJ	ˆK	ˆL	ˆM	ˆN	ˆO
0010	ˆP	ˆQ	ˆR	ˆS	ˆT	ˆU	ˆV	ˆW
0011	ˆX	ˆY	ˆZ	ˆ[	ˆ\	ˆ]	ˆˆ	ˆ_
0100		!	”	#	$	%	&	’
0101	(	)	*	+	,	-	.	/
0110	0	1	2	3	4	5	6	7
0111	8	9	:	;	<	=	>	?
1000	@	A	B	C	D	E	F	G
1001	H	I	J	K	L	M	N	O
1010	P	Q	R	S	T	U	V	W
1011	X	Y	Z	[	\	]	ˆ	_
1100	‘	a	b	c	d	e	f	g
1101	h	i	j	k	l	m	n	o
1110	p	q	r	s	t	u	v	w
1111	x	y	z	{	|	}	∼	DEL

[3] ASCII stands for "American Standard Code for Information InterchangeAmerican Standard Code for Information Interchange".

To use the table, find the letter you would like to encode. The first four binary digits are found in the leftmost column in the same row as the character and the final three, in the topmost row. For example, if you want to encode an upper-case "T" you will find 1010 to the left and 100 at the top, so the ASCII encoding for "T" is 1010100.

Exercise: Show that the ASCII encoding for the string "R2D2 dies" (don't omit the space!) is given by:

1010010 0110010 1000100 0110010 0100000 1100100 1101001 1100101 1110011

Since the recipient knows that every seven bits represents a character there's no reason not to pack them together to obtain the following 63-bit message:

101001001100101000100011001001000001100100110100111001011110011

In general, an n-bit message will require $7n$ bits to encode it in ASCII. In practice, you will send chunks of data, say 1000 bits at a time, so if you have a 10000-bit message, you would encode it as 10 1000-bit chunks.

There is nothing secret about ASCII: everyone in the world can find the table above and read it, so although it's sometimes called an "ASCII encoding" it is certainly not secret. What we need to do is to obfuscate the message before we transmit it.

9.2 The XOR Operation

The term "XOR" is the name of a logical binary operation that takes two bits and returns one bit. If we think of "1" as meaning "true" and "0" as meaning "false", then "XOR" stands for "exclusive OR". It is true if exactly one of the two incoming bits is true (1) and the other is false. If both are true or both are false, the XOR function returns false (0).

An equivalent way to think of XOR is that it is simply addition in base-2 where you throw away any carry bit. So $0 + 0 = 0, 0 + 1 = 1, 1 + 0 = 1$ and $1 + 1 = 0$.

Here is the XOR truth table:

A	B	(A XOR B)
0	0	0
0	1	1
1	0	1
1	1	0

Let's illustrate the encoding of a 2-letter message (we'll just use the first two letters, "R2" or 10100100110010, from the example in the exercise above. To encode it, we need a random 14-bit key, and assume that we've decided that 10111100101001 is a suitable key. To encode the "R2" message, we place the bits for "R2" above the bits for the key and form a third line where the i^{th} entry in the encoding is obtained by computing (A XOR B) where A and B are the i^{th} bits of the message and the key, respectively:

Message (M)	10100100110010
Key (K)	10111100101001
Cipher (M XOR K)	00011000011011

The first three bits in the message and the key are the same, so the first three bits of the cipher text are 0 since XOR returns 0 if the inputs are the same. The next two have differing inputs, so we get 1's in the cipher text, et cetera.

Exercise: Verify the remaining bits in the cipher text above.

The beauty of the XOR function is that if you XOR with the key again, you will obtain the original message. Check the result below and see that if you XOR the cipher text with the key, you'll be back to the orignal message:

Cipher (C)	00011000011011
Key (K)	10111100101001
(C XOR K)	10100100110010

Exercise: Show that if we assume that the n-bit key is chosen completely randomly (in other words, if every one of the 2^n possible n-bit keys is equally-likely) then every possible n-bit cipher is equally likely. In other words, if you have no idea what the random key was, the cipher message could equally-likely stand for any possible message.

If it is somehow possible to generate a truly-random key of length n and to securely get copies of that key to both the sender and recipient of the message with no chance of an opponent obtaining the key, then an n-bit message can be transmitted using the technique above with absolute security. Note, though, that you can't re-use the key, since if it is re-used, it begins to be possible to figure out something about the distribution of its bits.

9.3 Transmitting a Key

Here is the problem we would like to solve using EEC: The "good guys", Alice and Bob, wish to send and receive messages and to keep the contents of the messages secret from Eve (the "eavesdropper"). All of the transmissions between Alice and Bob occur on an insecure line which they know Eve has tapped, so Eve can see every message that Alice and Bob exchange.

Assume that there are no transmission errors so whenever Alice or Bob transmits a message, everyone winds up with a perfect copy (including Eve). Alice and Bob have good computers, but assume that Eve has a supercomputer that is 100 times as fast that the machines that Alice and Bob have. Also assume that Eve is twice as smart as Alice and Bob.

How do Alice and Bob communicate without letting Eve know any of their secrets?

The easiest solution is for Bob and Alice to get together before they need to make any transmissions and to make two copies of keys obtained by a random process, like radioactive decay or similar. This is a perfect solution, but if the agreed-upon data is n-bits long, they can only transmit n bits without having to get together again.

(It is also impractical in a real-world sense: quite often you will be "Alice" and "Bob" will be an internet store with a warehouse across the country, and you need to transmit in secret things like your credit card number. It is totally impractical for you to meet with "Bob" in person to share a secret key.)

So a harder problem is this: assume that Alice and Bob have to somehow agree on a secret key without any prior contact and so all the negotiation must take place over the insecure transmission line, so Eve can see the entire negotiation as well. Thus just sending the key is no good: Eve will have a copy of it as well and can decode the messages.

The method is actually quite simple (once you understand or at least believe all of the previous information in this article). Here's how to do it. Let's assume that Alice wants to send Bob a 300-bit message. If she wants to send more, just repeat the steps below as many times as necessary to transmit the entire message in 300-bit chunks.

1. Alice, using her computer, generates a 100-digit prime number p (which is more than 300 binary bits in base-2). She also generates a random point $P = (x, y)$ whose coordinates have about 300 bits each. Find a random A (also about 300 bits long). Since Alice knows x, y and A it is easy to find B so that $y^2 = x^3 + Ax + B$ – easy since $B = y^2 - x^3 - Ax \mod p$.

2. Alice transmits all of the above information: x, y, A, B and p, so she, Bob and Eve know all of it.

3. Alice invents another 300-bit number, call it a, but does *not* transmit it. Bob similarly invents a random 300-bit number that he calls b, but he does not transmit it, either.

4. Alice computes aP and transmits it. Bob computes bP and transmits that. At this point, everybody, including Eve, know: $P = (x, y), A, B, p, aP$ and bP. Note that since computing a multiple of P is a trapdoor operation that Eve can't easily compute a or b from aP or bP. But Alice knows a and Bob knows b.

5. Alice computes $a(bP) = (ab)P = K$ and Bob computes $b(aP) = (ba)P = (ab)P = K$. Both of them now know a pair of 300-bit coordinates that form a key K (we can just use the x-coordinate of K, if we like). Alice XORs the x-coordinate of K with her message and transmits it to Bob. Bob then XORs the transmitted message with his copy of the same x-coordinate of K, and reconstructs the message. Since the x and y coordinates are correlated by the equation of the curve, only one of them should be used as a key.

In the following section we will go through a toy example.

10 EEC Numerical Example

The numbers here are tiny compared to a real ECC implementation, but the operations can be checked with a hand calculator, if desired and it's not necessary to work with 100-digit primes or similar.

Suppose Alice wants to transmit a 7-bit secret message (a single ASCII character). Since $2^7 = 128$ we'll need a prime number p that's larger than 127; let's use $p = 149$. Pick a starting point P will coordinates between 0 and $148 = p - 1$, say $P = (x, y) = (23, 67)$. Pick $A = 111$ and we can calculate (all calculations are modulo 149) $B = y^2 - x^3 - Ax = 50$. Thus our elliptic equation is $y^2 = x^3 + 111x + 50$. Alice also checks that the curve is non-singular: $4A^3 + 27B^2 = 141$, so things are fine.

All the information above is transmitted so Alice, Bob and Eve know everything in the previous paragraph. But now Alice picks her private $a = 12$ and Bob picks a private $b = 77$. Alice computes $12P = (137, 62)$ and Bob computes $77P = (23, 82)$. Both $12P$ and $77P$ are transmitted (but not the 12 or the 77). Alice then computes $(12 \cdot 77P) = (137, 87)$ and Bob computes $(77 \cdot 12P) = (137, 87)$.

Alice uses 137, or 10001001 in binary, as her key to XOR with her secret message. Say the message is the character $X = 01011000$. 10001001 XOR 01011000 = 11010001 and Alice transmits that. Bob calculates 11010001 XOR 10001001 = 01011000 which he looks up in the ASCII table and determines that Alice's message was "X".

11 Practical Considerations

It is *not* a good idea to choose a random prime number and a random cubic equation as we did in the previous section. We'll use that example to show how things can go very wrong with arbitrary choices.

In that example, $p = 149$ and we used the equation $y^2 = x^3 + 111x + 50$ and $P = (23, 67)$ as our starting point. With such a small prime p we can exhaustively search for the number of solutions to the equation and we find that there are 156 of them, including $\mathcal{O}$, the point at infinity.

If we look at P, $2p$, $3P$, $\ldots$, we find that $39P = \mathcal{O}$ so as the multiples of P increase and increase, they cycle around every 39 times. This is much smaller (in fact, exactly $1/4$ as large as) the 156 total possibilities. Was $(23, 67)$ a bad choice? Well, yes. If we'd picked $(7, 53)$ as a starting point, all the points $P, 2P, 3P, \ldots, 156P$ are different, so the cycling would have taken the full 156 steps to return to the original value.

But although $(23, 67)$ only gave us 39 steps, we could have done much worse. What if our starting point had been, $78(7, 53) = (90, 0)$? We would find that $(90, 0) + (90, 0) = \mathcal{O}$ so it would alternate between $(90, 0)$ and $\mathcal{O}$ and the only thing that would matter about Alice's and Bob's choices for a and b would be whether they were even or odd. (The choice of 78 was made since it's half of 156.)

In any case, there are some subtleties involved in picking the prime, the associate equation, and the starting value. There are some choices that seem to work well in the "brainpool" data, available at:

`http://www.ecc-brainpool.org/download/Domain-parameters.pdf`

although who knows if their data has been hacked by the NSA?