Lecture 14: Elliptic Curve Cryptography

Lecture Notes on "Introduction to Computer Security"

by Avi Kak (kak@purdue.edu)

March 25, 2007

Goals:

- Introduction to elliptic curves
- A group structure imposed on the points on an elliptic curve
- Geometric and algebraic interpretations of the group operator
- Elliptic curves on prime finite fields
- Elliptic curves on Galois fields
- Elliptic curve cryptography
- Security of Elliptic Curve Cryptography

Why Elliptic Curve Cryptography?

- As you saw from the lecture on RSA, the computational overhead of that approach to public-key cryptography increases with the size of the keys. As algorithms for integer factorization have become more and more efficient, the RSA based methods have had to resort to longer keys.

- Elliptic curve cryptography can provide the same level and type of security as RSA (or Diffie-Hellman) **but with much shorter keys**.

- The table compares the key sizes for the different approaches to encryption for comparable levels of security against brute-force attacks. **What makes this table all the more significant is that for comparable key lengths the computational burdens of RSA and ECC are comparable.**

Symmetric Encryption Key Size in bits	*RSA and Diffie-Hellman Key size in bits*	*Elliptic Curve Key Size in bits*
80	1024	160
112	2048	224
128	3072	256
192	7680	384
256	15360	512

- The computational overhead of both RSA and ECC grows as $O(N^3)$ where N is the key length in bits. [Source: Hank van Tilborg, NAW, 2001]

- Another way to compare ECC with RSA is that the security of ECC grows exponentially in its parameters, whereas the security of RSA grows only subexponentially in its parameters.

- Because of the much smaller key sizes involved, ECC algorithms can be implemented on **smartcards** without mathematical co-processors. **Contactless smart cards** work only with ECC because other systems require too much induction energy. Since shorter key lengths translate into faster handshaking protocols, ECC is also becoming increasingly important for **wireless communications**. [Source: Hank van Tilborg, NAW, 2001]

- For the same reasons as listed above, we can also expect ECC to become important for **wireless sensor networks**.

What are Elliptic Curves?

- First and foremost, elliptic curves have nothing to do with ellipses. Ellipses are formed by quadratic curves. Elliptic curves are always cubic. [Note: Elliptic curves are called **elliptic** because of their relationship to **elliptic integrals** in mathematics. An elliptic integral can be used to determine the arc length of an ellipse.]

- The simplest possible "curves" are, of course, straight lines.

- The next simplest possible curves are conics, these being quadratic forms of the following sort

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

If $b^2 - 4ac$ is less than 0, then the curve is either an ellipse, or a circle, or a point, or the curve does not exist; if it is equal to 0, then we have either a parabola, or two parallel lines, or no curve at all; if it is greater than 0, then we either have a hyperbola or two intersecting lines. (Note that, by definition, a conic is the intersection of a plane and a cone.)

- The next simplest possible curves are elliptic curves. An elliptic curve in its "standard form" is described by

$$y^2 = x^3 + ax + b$$

for some fixed values for the parameters a and b. This equation is also referred to as **Weierstrass Equation** of **characteristic** 0. (The equation shown involves multiplications and additions over certain objects that are represented by x, y, a, and b. The values that these object acquire are meant to be drawn from a set that must at least be a **ring**. The **characteristic** of a ring is the number of times you must add the multiplicative identity element in order to get the additive identity element. If adding the multiplicative identity element to itself, no matter how many times, **never** gives us the additive identity element, we say the characteristic is 0. Otherwise, there must exist an integer p such that $p \times n = 0$ for all n. The value of p is then the characteristic of the ring. In a **ring of characteristic** 2, the elements 2, 4, etc., are all equal to 0. In a **ring of characteristic** 3, the elements 3, 6, etc., are all equal to 0.) Elliptic curves have a rich structure that can be put to use for cryptography.

- Slide 7 shows some elliptic curves for a set of parameters $(a,\ b)$. The top four curves all look smooth (they do not have cusps, for example) because they all satisfy the following condition on the **discriminant** of the polynomial $f(x) = x^3 + ax + b$:

$$4a^3 + 27b^2 \neq 0 \tag{1}$$

[Note: The discriminant of a polynomial is the product of the squares of the differences of the polynomial roots. The roots of the polynomial $f(x) = x^3 + ax + b$ are obtained by solving the equation $x^3 + ax + b = 0$. Since this is a cubic polynomial, it will in general have three roots. Let's call them r_1, r_2, and r_3. Its discriminant will therefore be

$$D_3 = \prod_{i<j}^{3} (r_i - r_j)^2$$

which is the same as $(r_1 - r_2)^2(r_1 - r_3)^2(r_2 - r_3)^2$. It can be shown that when the

polynomial is $x^3 + ax + b$, the discriminant reduces to

$$D_3 \quad = \quad -16(4a^3 \quad + \quad 27b^2)$$

This discriminant must not become zero for an elliptic curve polynomial $x^3 + ax + b$ to possess three distinct roots. If the discriminant is zero, that would imply that two or more roots have coalesced, giving the curve a cusp or some other form of non-smoothness. Non-smooth curves are **singular**. It is **not safe** to use singular curves for cryptography.]

- The **bottom two** examples on Slide 7 show two elliptic curves for which the condition on the discriminant is violated. For the one on the left that corresponds to $f(x) = x^3$, all three roots of the cubic polynomial have coalesced into a single point and we get a cusp at that point. For the one on the right that corresponds to $f(x) = x^3 - 3x + 2$, two of the roots have coalesced into the point where the curve crosses itself. These two curves are **singular**. As mentioned earlier, it is **not safe** to use singular curves for cryptography.

- Note that since we can write

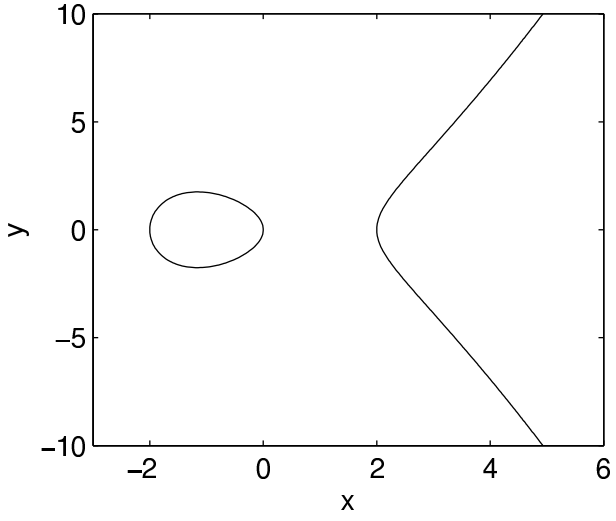
$$y \quad = \quad \pm\sqrt{x^3 \quad + \quad ax \quad + \quad b}$$

elliptic curves in their standard form will be symmetric about the x-axis.

- It is difficult to comprehend the structure of the curves that involve polynomials of degree greater than 3.

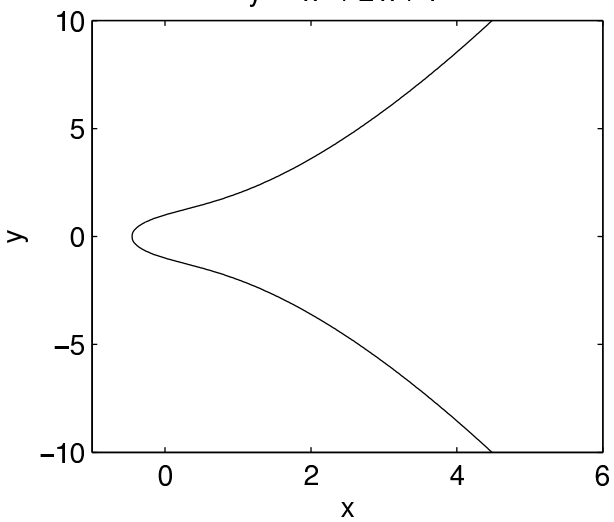
$4a^3 + 27b^2 < 0$

$y^2 = x^3 - 4x$

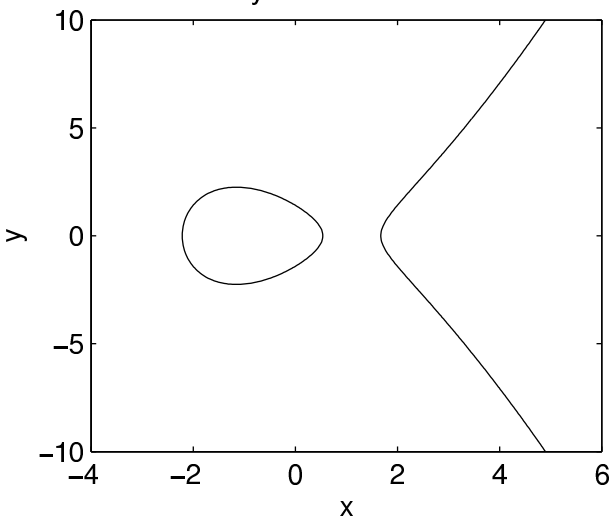


$4a^3 + 27b^2 > 0$

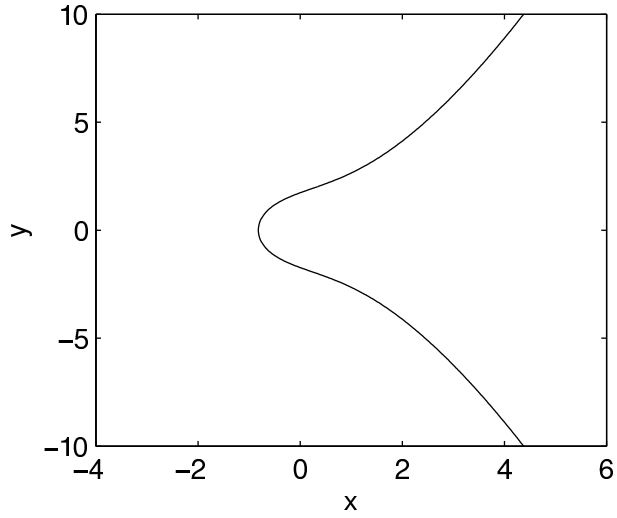
$y^2 = x^3 + 2x + 1$



$y^2 = x^3 - 4x + 2$

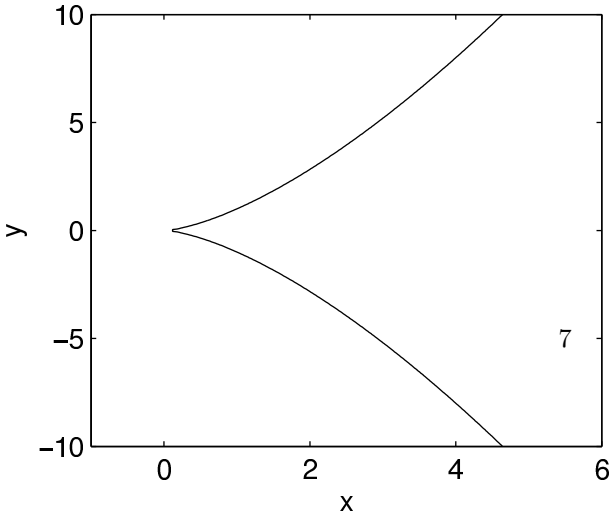


$y^2 = x^3 + 3x + 3$

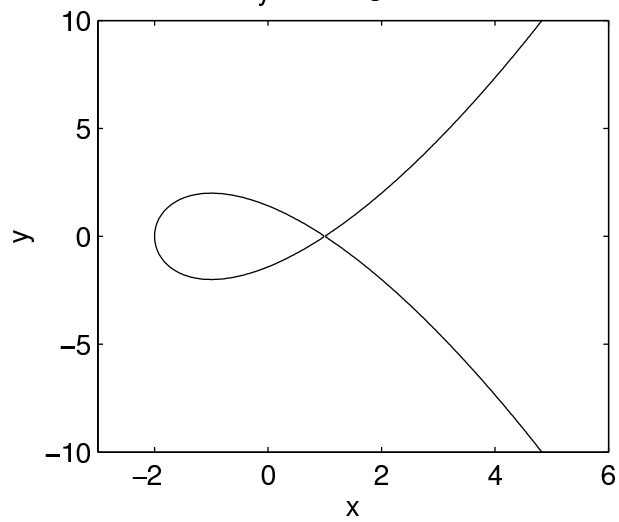


Condition on the discriminant violated in the figures below

$y^2 = x^3$



$y^2 = x^3 - 3x + 2$



A Group Law Defined for Points on an Elliptic Curve

- The points on an elliptic curve can be shown to constitute a group.

- Recall that a group needs: a group operator; an identity element with respect to the operator; closure and associativity with respect to the operator; and the existence of inverses with respect to the operator.

- The group operator for the points on an elliptic curve is, by convention, called **addition**. Its definition has nothing to do with the conventional arithmetic addition.

- To add a point P on an elliptic curve to another point Q on the same curve, we use the following rule

  - We first join P with Q with a straight line. The third point of the intersection of this straight line with the curve, if such an intersection exists, is denoted R. The mirror image of this point with respect to the x-coordinate is the point $P + Q$. If the third point of intersection does **not** exist, we say it is at **infinity**.

– The addition operation is shown on Slide 13 for two different elliptic curves. The values for a and b for the curve at the top are -1 and 0, and for the curve at the bottom 1 and 1.

- But what happens when the intersection of P and Q is at infinity?

- We denote the point at infinity by the special symbol $\mathbf{O}$ and we then show that this can serve as the additive identity element for the group operator.

- We now stipulate that that $P + \mathbf{O} = P$ for any point on the curve.

- We define the additive inverse of a point P as its mirror reflection with respect to the x coordinate. So if Q on the curve is the mirror reflection of P on the curve, then $Q = -P$. For any such two points, it would obviously be the case that the third point of intersection will be at infinity. That is, the third point of intersection will be the distinguished point $\mathbf{O}$.

- We will further stipulate that that $\mathbf{O} + \mathbf{O} = \mathbf{O}$, implying that $-\mathbf{O} = \mathbf{O}$.

- Therefore, the mirror reflection of the point at infinity is the same point at infinity.

- Now we can go back to the issue of what happens to $P + Q$ when the intersection of two points P and Q is at infinity. Obviously, in this case, the intersection is at the distinguished point $\mathbf{O}$, whose additive inverse is also $\mathbf{O}$. Therefore, for such points, $P + Q = \mathbf{O}$.

- We have already defined the additive inverse of a point P as its mirror reflection about the x-axis. What is the additive inverse of a point where the tangent is parallel to the y-axis? The additive inverse of such a point is the point itself. That is, if the tangent at P is parallel to the y-axis, then $P + P = \mathbf{O}$.

- In general, what does it mean to add P to itself? To see what it means, let's consider two distinct points P and Q and let Q approach P. The line joining P and Q will obviously become a tangent at P in the limit. Therefore, the operation $P + P$ means that we must draw a tangent at P, find the intersection of the tangent with the curve, and then take the mirror reflection of the intersection.

- Obviously, if the tangent at P intersects the curve at infinity, meaning at the distinguished point $\mathbf{O}$, then $P + P = \mathbf{O}$.

- For an elliptic curve

$$y^2 \quad = \quad x^3 \quad + \quad ax \quad + \quad b$$

we define the set of all points on the curve along with the distinguished point $\mathbf{O}$ by $E(a, b)$.

- $E(a, b)$ is a group with the "addition" operator as defined on Slides 8, 9, and 10.

- $E(a, b)$ is obviously closed with respect to the addition operation. We can also show geometrically that the property of associativity is satisfied. Every element in the set obviously has its additive inverse in the set.

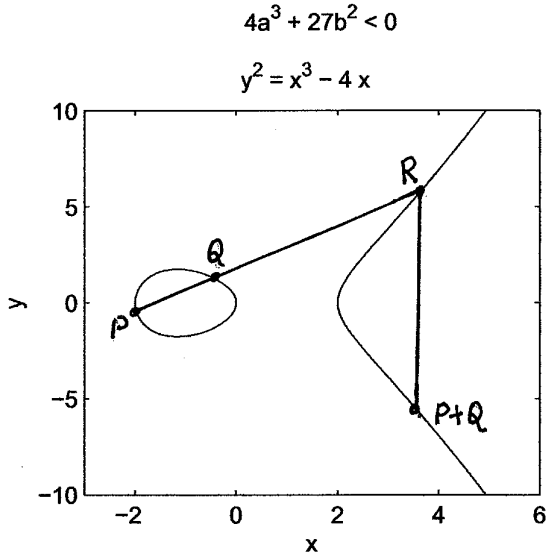
- Since the operation of "addition" is commutative, $E(a, b)$ is an **abelian group**.

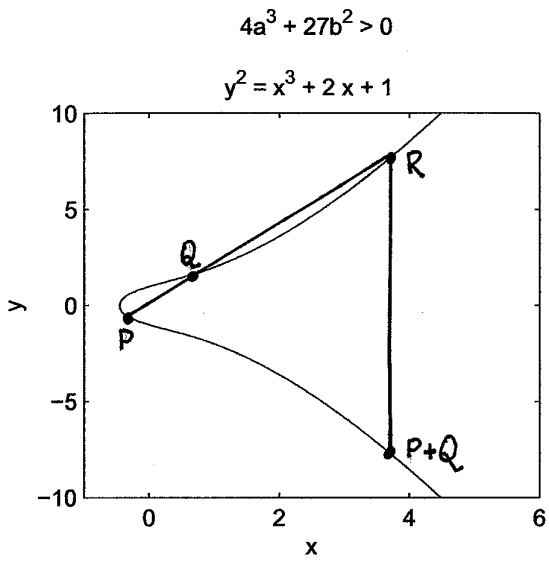
- Just for notational convenience, we now define multiplication on this group as repeated addition. Therefore,

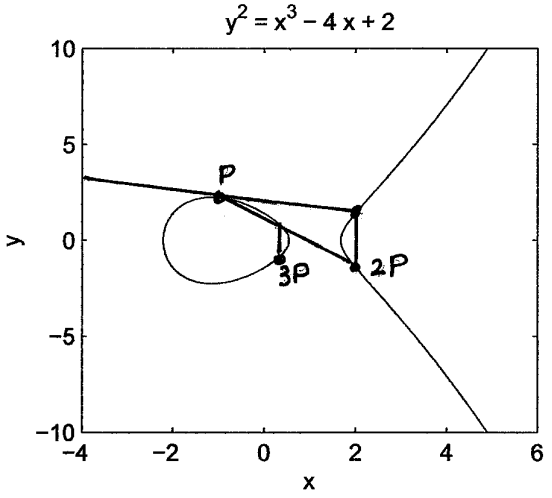
$$k \times P \quad = \quad P \quad + \quad P \quad + \quad \ldots \quad + \quad P$$

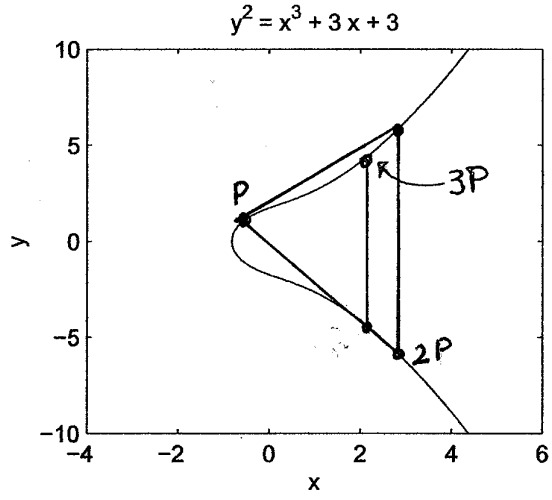
with P making k appearances on the right.

- Therefore, we can express $P \ + \ P$ as $2P$, $P \ + \ P \ + \ P$ as $3P$, and so on.


$4a^3 + 27b^2 < 0$
$y^2 = x^3 - 4x$
P
Q
R
P+Q
x
y


$4a^3 + 27b^2 > 0$
$y^2 = x^3 + 2x + 1$
P
Q
R
P+Q
x
y


$y^2 = x^3 - 4x + 2$
P
3P
2P
x
y


$y^2 = x^3 + 3x + 3$
P
3P
2P
x
y

The Characteristic of the Underlying Field and the Singular Elliptic Curves

- The examples of the elliptic curves shown so far were for **the field of real numbers**. These fields are of characteristic zero because no matter how many multiplicative identity elements you add, you'll never get an additive identity element.

- The group law can also be defined when the underlying field is characteristic 2 or 3. But now the elliptic curve $y^2 = x^3 + ax + b$ becomes **singular**. While singular elliptic curves do admit group laws of the sort we showed on the previous slides, such groups become **isomorphic** to either the multiplicative or the additive group over the underlying field, depending on the type of singularity. **That fact makes singular elliptic curves unsuitable for cryptography because they are easy to crack.**

- To show that the elliptic curve $y^2 = x^3 + ax + b$ becomes singular when the characteristic of the underlying field is 2, let's look at the partial derivatives of the two sides of the equation of this curve:

$$2ydy \;\; = \;\; 3x^2dx \; + \; adx$$

  implying

$$\frac{dy}{dx} = \frac{3x^2 + a}{2y} \tag{2}$$

- A point on the curve is **singular** if $\frac{dy}{ds}$ is not well defined. This would be the point where both the numerator and the denominator are zero. [The slope is defined, even though it is ∞, when only the denominator goes to zero.] So the elliptic curve $y^2 = x^3 + ax + b$ will become singular if it contains a point (x, y) so that

$$\begin{aligned} 3x^2 + a &= 0 \\ 2y &= 0 \end{aligned}$$

  and the point (x, y) satisfying these two equations lies on the curve.

- Let's now consider the case when the underlying field is of characteristic 2. In this case, we go back to Equation (2) above and see that, since 2 is the same thing as 0 for such a field, the derivative $\frac{dy}{dx}$ will **not** be defined at $x = \sqrt{\frac{-a}{3}}$. Therefore, the curve $y^2 = x^3 + ax + b$ will be singular for some values of a that can be obtained by substituting $x = \sqrt{\frac{-a}{3}}$ in the equation of the curve.

- Let's now consider the case of a field of characteristic 3. In this case, since 3 is the same thing as 0, we can write for the curve slope from Equation (2):

$$\frac{dy}{dx} = \frac{a}{2y}$$

This curve becomes singular if we should choose $a = 0$.

- In general, when using the elliptic curve equation $y^2 = x^3+ax+b$, we avoid underlying fields of characteristic 2 or 3 because of the nature of the constraints they place on the parameters a and b in order for the curve to not become singular.

An Algebraic Expression for Adding Two Points on An Elliptic Curve

- Given two points P and Q on an elliptic curve $E(a, b)$, we have already pointed out that to compute the point $P + Q$, we first draw a straight line through P and Q. We next find the third intersection of this line with the elliptic curve. We denote this point of intersection by R. Then $P + Q$ is equal to the mirror reflection of R about the x-axis.

- In other words, if P, Q, and R are the three intersections of the straight line with the curve, then

$$P + Q = -R$$

- This implies that the three intersections of a straight line with the elliptic curve must satisfy

$$P + Q + R = \mathbf{O}$$

- We will next examine the algebraic implications of the above relationship between the three points of intersection.

- The equation of the straight line that runs through the points P and Q is obviously of the form:

$$y \quad = \quad \alpha x \quad + \quad \beta$$

where α is the slope of the line, which is given by

$$\alpha \quad = \quad \frac{y_Q \; - \; y_P}{x_Q \; - \; x_P}$$

- For a point $(x,\ y)$ to lie at the intersection of the straight line and the elliptic curve $E(a, b)$, the following equality must obviously hold

$$(\alpha x \; + \; \beta)^2 \quad = \quad x^3 \quad + \quad ax \quad + \quad b \qquad (3)$$

since $y \; = \; \alpha x \; + \; \beta$ on the straight line through the points P and Q and since the equation of the elliptic curve is $y^2 \; = \; x^3 + ax + b$.

- For there to be three points of intersection between the straight line and the elliptic curve, the cubic form in Equation (3) must obviously have three roots. **We already know two of these roots, since they must be x_P and x_Q, correspond to the points P and Q.**

- Being a cubic equation, since Equation (3) has at most three roots, the remaining root must be x_R, the x-coordinate of the third point R.

- Equation (3) represents a **monic polynomial**. What that means is that the coefficient of the highest power of x is 1.

- **A property of monic polynomials is that the sum of their roots is equal to minus the coefficient of the second highest power.** Expressing Equation (3) in the following form:

$$x^3 - \alpha^2 x^2 + (a - 2\alpha\beta)x + (b - \beta^2) = 0 \quad (4)$$

we notice that the coefficient of x^2 is $-\alpha^2$. Therefore, we have

$$x_P + x_Q + x_R = \alpha^2$$

We therefore have the following result for the x-coordinate of R:

$$x_R = \alpha^2 - x_P - x_Q \quad (5)$$

- Since the point (x_R, y_R) must be on the straight line $y = \alpha x + \beta$, we can write for y_R:

$$\begin{aligned} y_R &= \alpha x_R + \beta \\ &= \alpha x_R + (y_P - \alpha x_P) \\ &= \alpha(x_R - x_P) + y_P \end{aligned} \tag{6}$$

- To summarize, ordinarily a straight line will intersect an elliptical curve at three points. If the coordinates of the first two points are (x_P, y_P) and (x_Q, y_Q), then the coordinates of the third point are

$$x_R = \alpha^2 - x_P - x_Q \tag{7}$$
$$y_R = \alpha(x_R - x_P) + y_P \tag{8}$$

- We started out with the following relationship between P, Q, and R

$$P + Q = -R$$

we can therefore write the following expressions for the x and the y coordinates of the addition of two points P and Q:

$$x_{P+Q} = \alpha^2 - x_P - x_Q \tag{9}$$
$$y_{P+Q} = -y_P + \alpha(x_P - x_R) \tag{10}$$

since the y-coordinate of the reflection $-R$ is negative of the y-coordinate of the point R on the intersecting straight line.

An Algebraic Expression for Calculating $2P$ from P

- Given a point P on the elliptical curve $E(a, b)$, computing $2P$ (which is the same thing as computing $P + P$), requires us to draw a tangent at P and to find the intersection of this tangent with the curve. The reflection of this intersection about the x-axis is then the value of $2P$.

- Given the equation of the elliptical curve $y^2 = x^3 + ax + b$, the slope of the tangent at a point (x, y) is obtained by differentiating both sides of the curve equation

$$2y\frac{dy}{dx} = 3x^2 + a$$

- We can therefore write the following expression for the slope of the tangent at point P:

$$\alpha = \frac{3x_P^2 + a}{2y_P} \tag{11}$$

- Since drawing the tangent at P is the limiting case of drawing a line through P and Q as Q approaches P, two of the three roots of the following equation (which is the same as Equation (3) you saw before):

$$(\alpha x + \beta)^2 = x^3 + ax + b \quad (12)$$

must coalesce into the point x_P and the third root must be x_R. As before, R is the point of intersection of the tangent with the elliptical curve.

- As before, we can use the property that sum of the roots of the monic polynomial above must equal the negative of the coefficient of the second highest power. Noting two of the three roots have coalesced into x_P, we get

$$x_P + x_P + x_R = \alpha^2$$

- Substituting the value of α from Equation (10) in the above equation, we get

$$x_R = \alpha^2 - 2x_P = \left(\frac{3x_P^2 + a}{2y_P}\right)^2 - 2x_P \quad (13)$$

- Since the point R must also lie on the straight line $y = \alpha x + \beta$, substituting the expression for x_R in this equation yields

$$\begin{aligned} y_R &= \alpha x_R + \beta \\ &= \alpha x_R + (y_P - \alpha x_P) \end{aligned}$$

$$
\begin{aligned}
&= \quad \alpha(x_R \;-\; x_P) \;+\; y_P \\
&= \quad \frac{3x_P^2 + a}{2y_P}(x_R \;-\; x_P) \;+\; y_P \qquad (14)
\end{aligned}
$$

- To summarize, if we draw a tangent at point P to an elliptical curve, the tangent will intersect the curve at a point R whose coordinates are given by

$$
\begin{aligned}
x_R &= \frac{3x_P^2 \;+\; a}{2y_P} \;-\; 2x_P \\
y_R &= \frac{3x_P^2 + a}{2y_P}(x_R \;-\; x_P) \;+\; y_P \qquad (15)
\end{aligned}
$$

- Since the value of $2P$ is the reflection of the point R about the x-axis, the value of $2P$ is obtained by taking the negative of the y-coordinate:

$$
\begin{aligned}
x_{2P} &= \frac{3x_P^2 \;+\; a}{2y_P} \;-\; 2x_P \\
y_{2P} &= \frac{3x_P^2 + a}{2y_P}(x_P \;-\; x_R) \;-\; y_P \qquad (16)
\end{aligned}
$$

Elliptic Curves Over Z_p for Prime p

- The elliptic curve arithmetic we described so far was over **real numbers**. These curves cannot be used as such for cryptography because calculations with real numbers are prone to round-off error. **Cryptography requires error-free arithmetic.**

- However, by restricting the values of the parameters a and b, the value of the independent variable x, and the value of the dependent variable y to belong to the **prime finite field** Z_p, we obtain elliptic curves that are more appropriate for cryptography:

$$y^2 \bmod p \quad = \quad (x^3 \quad + \quad ax \quad + \quad b) \bmod p \qquad (17)$$

  subject to the modulo p version of the same smoothness constraint on the discriminant as we had for the case of real numbers [see Equation (1) on Slide 5]:

$$(4a^3 \quad + \quad 27b^2) \bmod p \quad \neq \quad 0 \bmod p$$

- We will use the notation $E_p(a, b)$ to represent all the points (x, y) that obey the above equation. $E_p(a, b)$ will also include the distinguished point $\mathbf{O}$, the point at infinity.

- So the points in $E_p(a, b)$ are the set of coordinates (x, y), with $x, y \in Z_p$, such that the equation $y^2 = x^3 + ax + b$, with $a, b \in Z_p$ is satisfied modulo p and such that the condition $4a^3 + 27b^2 \neq 0 \pmod{p}$ is fulfilled.

- Obviously, then, the set of points in $E_p(a, b)$ is no longer a curve, but a collection of discrete points in the (x, y) plane (or, even more precisely speaking, in the plane corresponding to the Cartesian product $Z_p \times Z_p$).

- Since the points in $E_p(a, b)$ can no longer be connected to form a smooth curve, we cannot use the geometrical construction to illustrate the action of the group operator. That is, given a point P, now one cannot show geometrically how to compute $2P$, or given two points P and Q, one cannot show geometrically how to determine $P + Q$. **However, the algebraic expressions we derived for these operations continue to hold good provided the calculations are carried out modulo p.**

- Note that for a **prime finite field** Z_p, the value of p is its **characteristic**. (See Slide 5 for what is meant by the characteristic of a ring.) Elliptic curves over **prime finite fields** with $p \leq 3$, while admitting the group law, are **not** suitable for cryptography. (See Slide 14)

- As we will see on the next slide, elliptic curves can also be defined over Galois Fields $GF(2^m)$. (These are also commonly denoted Z_{2^m} and commonly called **binary finite fields**.) Binary finite fields have characteristic 2.

Elliptic Curves Over Galois Fields $GF(2^m)$

- For hardware implementations of ECC, it is common to define elliptic curves over a Galois Field $GF(2^n)$. (In contrast with the **primary finite fields** denoted by Z_p for prime p, Galois Fields are also called **binary finite fields**.)

- What makes the binary finite fields more convenient for hardware implementations is that the elements of $GF(2^n)$ can be represented by n-bit binary code words. (See Part 4 of the lecture slides on Finite Fields.)

- You will recall from the lecture on finite fields, that the addition operation in $GF(2^n)$ is like the XOR operation on bit fields. That is $x + x = 0$ for all $x \in GF(2^n)$. This implies that a finite field of form $GF(2^n)$ is of **characteristic** 2. (See Slide 5 for what is meant by the **characteristic** of a finite field.)

- As mentioned before, the elliptic curve we showed earlier ($y^2 = x^3 + ax + b$) is meant to be used only when the underlying finite field is of characteristic greater than 3. (See Slide 14)

- The elliptic curve equation to use when the underlying field is described by $GF(2^n)$ is

$$y^2 \;+\; xy \;\;=\;\; x^3 \;+\; ax^2 \;+b, \qquad b \neq 0 \qquad (18)$$

  The constraint $b \neq 0$ serves the same purpose here that the constraint $4a^3 + 27b^2 \neq 0$ did for the case of the elliptic curve equation $y^2 = x^3 + ax + b$. The reason for the constraint $b \neq 0$ is that the discriminant becomes 0 when $b = 0$. As mentioned earlier, when the discriminant becomes zero, we have multiple roots at the same point, causing the derivative of the curve to become ill-defined at that point. In other words, the curve has a singularity at the point where discriminant is 0.

- The fact that the equation of the elliptic curve is different when the underlying field is $GF(2^n)$ introduces the following changes in the behavior of the group operator:

  - Given a point $P = (x,\ y)$, the negative of this point would be located at $-P = (x,\ x + y)$.

  - Given two distinct points $P = (x_P,\ y_P)$ and $Q = (x_Q,\ y_Q)$, the addition of the two points, represented by $(x_{P+Q},\ y_{P+Q})$, is given by

$$x_{P+Q} \;\;=\;\; \lambda^2 \;+\; \lambda \;+\; x_P \;+\; x_Q \;+\; a$$

$$y_{P+Q} \quad = \quad \lambda(x_P \ + \ x_{P+Q}) \ \ + \ \ x_{P+Q} \ \ + \ \ y_P$$

with

$$\lambda \quad = \quad \frac{y_P \ + \ y_Q}{x_P \ + \ x_Q}$$

– To double a point, that is to calculate $2P$ from P, we use the formulas

$$\begin{aligned} x_{2P} \quad &= \quad \lambda^2 \ + \ \lambda \ + \ a \\ y_{2P} \quad &= \quad x_P^2 \ + \ (\lambda \ + \ 1)x_{2P} \end{aligned}$$

with

$$\lambda \quad = \quad x_P \ + \ \frac{y_P}{x_P}$$

- We will use the notation $E_{2^n}(a, b)$ to denote the set of all points $(x, y) \in GF(2^n) \times GF(2^n)$, that satisfy the equation

$$y^2 \ + \ xy \quad = \quad x^3 \ + \ ax^2 \ + \ b,$$

with $a \in GF(2^n)$ and $b \in GF(2^n)$, along with the distinguished point $\mathbf{O}$ that serves as the additive identity element for the group structure formed by the points on the curve. Note that we do not allow b in the above equation to take on the value which is the additive identity element of the finite field $GF(2^n)$.

- If g is a generator for the field $GF(2^n)$ (see the Lecture notes on finite fields for what is meant by a generator here), then all the element of $GF(2^n)$ can be expressed in the following form

$$0,\ 1,\ g,\ g,\ g^2,\ g^3, ..., g^{2^n-2}$$

This implies that the majority of the points on the elliptic curve $E_{2^n}(a, b)$ can be expressed in the form $(g^i,\ g^j)$, where $i, j = 0,\ 1,\ \ldots,\ n-2$. In addition, there may be points whose coordinates can be expressed $(0, g^i)$ or $(g^i, 0)$, with $i = 0,\ 1,\ \ldots,\ n-2$. And then there is, of course, the distinguished point **O**.

- The **order of an elliptic curve**, that is the number of point in the group $E_{2^n}(a, b)$, in relation to the number of elements in $GF(2^n)$ **is important from the standpoint of the cryptographic security of the curve.** [Note: When we talk about the order of $E_{2^n}(a, b)$, we must of course include the distinguished point **O**.]

- Hasse's Theorem addresses the question of how many points are on an elliptic curve that is defined over a **finite** field. This theorem says that if N is the number of points on $E_q(a, b)$ when the curve is defined on a finite field Z_q with q elements, then N is bounded by

$$|N\ -\ (q\ +\ 1)| \quad \leq \quad 2\sqrt{q}$$

As mentioned previously, N includes the additive identity element **O**.

- Since the Galois field $GF(2^n)$ contains 2^n elements, we can say that the **order** of $E_{2^n}(a, b)$ is equal to $2^n + 1 - t$ where t is a number such that $|t| \leq \sqrt{2^n}$.

- An elliptic curve defined over a Galois Field $GF(2^n)$ is **supersingular** if $2|t$, that is if 2 is a divisor of t. [Supersingularity is **not** to be confused with singularity. When an elliptic curve is defined over real numbers, singularity of the curve is related to its smoothness. More specifically, a curve is singular if its slope at a point is not defined. **Supersingularity**, on the other hand, is related to the order of E_{2^n} and how this order relates to the number of points in the underlying finite field.]

- Should it happen that $t = 0$, then the order of E_{2^n} is $2n + 1$. Since this number is always odd, such a curve can never be supersingular. Supersingular curves defined over fields of characteristic 2 (which includes the binary finite fields $GF(2^n)$) always have an odd number of points, including the distingushed point $\mathbf{O}$.

- Supersingular curves are to be avoided for cryptography because they are vulnerable to the MOV attack. More on that later.

Is $b \neq 0$ a Sufficient Condition for the Elliptic Curve $y^2 + xy = x^3 + ax^2 + b$ to Not Be Singular?

- In general, we want to avoid using **singular** elliptic curves for cryptography for reasons already indicated.

- On Slide 28 we indicated that when using a curve of form $y^2 + xy = x^3 + ax^2 + b$, you want to make sure that $b \neq 0$ since otherwise the curve will be singular.

- We will now consider in greater detail when exactly the curve $y^2 + xy = x^3 + ax^2 + b$ becomes singular for the case when the underlying field consists of real numbers. Toward that end we will derive an expression for the discriminant of a polynomial that is singular if and only if the curve $y^2+xy = x^3+ax^2+b$ is singular. The condition which will prevent the discriminant going to zero will be the condition under which the curve $y^2+xy = x^3+ax^2+b$ will stay nonsingular.

- To meet the goal stated above, we will introduce the coordinate transformation

$$y \quad = \quad Y \; - \; \frac{x}{2}$$

in the equation

$$y^2 \; + \; xy \quad = \quad x^3 \; + \; ax^2 \; + b$$

- The purpose of the coordinate transformation is to get rid of the troublesome term xy in the equation. Note that this coordinate transformation cannot make a singularity disappear, and neither can it introduce a new singularity. With this transformation, the equation of the curve becomes

$$Y^2 - \frac{x^2}{4} = x^3 + ax^2 + b$$

which can be rewritten as

$$Y^2 = x^3 + (a + \frac{1}{4})x^2 + b$$

The polynomial on the right hand side of the equation shown above has a singular point wherever its discriminant goes to zero.

- In general, the discriminant of the polynomial

$$a_3 z^3 + a_2 z^2 + a_1 z = 0$$

is given by

$$D_3 = a_1^2 a_2^2 - 4a_0 a_2^3 - 4a_1^3 a_3 + 18 a_0 a_1 a_2 a_3 - 27 a_0^2 a_3^2$$

- Substituting the coefficient values for our case, $a_3 = 1$, $a_2 = (a + \frac{1}{4})$, $a_1 = 0$, and $a_0 = b$, in the general formula for the discriminant of a cubic polynomial, we get for the discriminant

$$D_3 = -4b\left(a + \frac{1}{4}\right)^3 - 27b^2$$

This simplifies to

$$D_3 = \frac{1}{16}\left[-64a^3 b - 48a^2 b - 12ab - b - 432b^2\right]$$

which can be expressed as

$$D_3 \quad = \quad -\frac{1}{16}b\left[64a^3 \ + \ 48a^2 \ + \ 12a \ + \ 432b \ + \ 1\right]$$

- Obviously, if $b = 0$, the discriminant will become 0. However, it is also obvious that even when the $b = 0$ condition is satisfied, certain values of a and b may cause the discriminant to go to 0.

- As with the supersingular curves, elliptic curves that are singular are to be avoided for cryptography because they are vulnerable to the MOV attack.

Elliptic Curve Cryptography

- That elliptic curves over finite fields could be used for cryptography was suggested independently by Neal Koblitz (University of Washington) and Victor Miller (IBM) in 1985.

- Just as RSA uses multiplication as its basic arithmetic operation (exponentiation is merely repeated multiplication), ECC uses the "addition" group operator as its basic arithmetic operation (multiplication is merely repeated addition).

- Suppose G is a user-chosen "base point" on the curve $E_q(a, b)$, where $q = p$ for some prime p when the underlying finite field is a prime finite field and $q = 2^n$ when the underlying finite field is a Galois field.

- In accordance with how the group operator works, $k \times G$ stands for $G + G + G + \ldots + G$ with G making k appearances in this expression.

- Now suppose our message consists of an integer M and we encrypt it by calculating $C = M \times G$. Now the question is whether an adversary with knowledge of all of the parameters of the curve $E_q(a, b)$ and of the point G can decrypt C.

- The core notion that ECC is based on is that, with a proper choice for G, whereas it is relatively easy to calculate $C = M \times G$, it can be extremely to recover M from C even when an adversary knows the curve $E_q(a, b)$ and the G used. Recovering M from C is referred to as having to solve the **discrete logarithm** problem. [To understand why finding M from C is referred to as solving the discrete logarithm problem: Note that word "addition" for the group operator for $E_q(a, b)$ is a matter of convention and convenience. As you already know from the lectures on finite fields, a group operator is typically referred to as addition and denoted '+', whereas the second operator when the group becomes a ring is typically called multiplication and denoted '×'. So there is nothing wrong with choosing to express $G + G + G + \ldots + G$ more generically as $G \circ G \circ G \circ \ldots \circ G$ if we do not want to get confused by mental associations with the '+' operator. **Now let's see what we mean by a logarithm.** As you know, if $a = b^n$ then $n = \log_b a$. We are at a liberty to write b^n as $b \times b \times b \ldots \times b$, or even as $b \circ b \circ b \ldots \circ b$ if we assume that the operator $\circ$ stands for multiplication. If we want to recover the **number of times** b participates in $a = b \circ b \circ b \ldots \circ b$ we take the logarithm of a to the base b. By the same token, if we want to determine the **number of times** G participates in $C = G \circ G \circ G \circ \ldots \circ G$, we take the "logarithm" of C to the base G.]

- An adversary could try to recover M from $C = MG$ by calculating $2G$, $3G$, $4G$, ..., kG with k spanning the size of the set $E_q(a, b)$, and then seeing which one of the results matched C. But if q is sufficiently large and if the point G on the curve $E_q(a, b)$ is chosen carefully, that would take much too long.

Elliptic Curve Diffie-Hellman Secret Key Exchange

- A community of users wishing to engage in secure communications with ECC chooses the parameters q, a, and b for an elliptic curve based group $E_q(a, b)$, and a base point $G \in E_q(a, b)$.

- A selects an integer PR_A to serve as his/her private key. A then generates $PU_A = PR_A \times G$ to serve as his/her public key. A makes publicly available the public key PU_A.

- B designates an integer PR_B to serve as his/her private key. As was done by A, B also calculates his/her public key by $PU_B = PR_B \times G$.

- In order to create a shared secret key (that could subsequently be used for, say, a symmetric-key based communication link), both A and B now carry out the following operations:

  - A calculates the shared secret key by

$$K = PR_A \times PU_B \tag{19}$$

– B calculates the shared secret key by

$$K \quad = \quad PR_B \times PU_A \tag{20}$$

– The calculations in Eqs. (17) and (18) yield the same result because

$$\begin{aligned} K \text{ as calculated by } A \quad &= \quad PR_A \times PU_B \\ &= \quad PR_A \times (PR_B \times G) \\ &= \quad (PR_A \times PR_B) \times G \\ &= \quad (PR_B \times PR_A) \times G \\ &= \quad PR_B \times (PR_A \times G) \\ &= \quad PR_B \times PU_A \\ &= \quad K \text{ as calculated by } B \end{aligned}$$

- To discover the secret key, an attacker could try to discover PR_A from the publicly available base point G and the publicly available PU_A. Recall, $PU_A = PR_A \times G$. **But this requires solving the discrete logarithm problem which, for a properly chosen set of curve parameters and G, can be extremely hard.**

- To increase the level of difficulty in solving the discrete logarithm problem, we select for G a base point whose **order** is very large. The **order** of a point on the elliptic curve is the **least number of times** G must be added to itself so that we get the **identity**

element 0 of the group $E_q(a, b)$. [We can also associate the notion of **order** with an elliptic curve over a finite field. The **order of an elliptic curve** is the total number of points in the set $E_q(a, b)$. This order is denoted $\#E_q(a, b)$.]

- Since the integers PR_A, PU_A, PR_B, and PU_B must all be less than the order n of the base point G, the value of n is also a part of the information that must be made publicly available.

- The base point G is also known as the **generator** of a **subgroup** of $E_q(a, b)$ whose elements are all given by G, $2G$, $3G$, ..., and, of course, the identity element **O**. For the size of the subgroup to equal the **degree** of the generator G, the value of n must be a prime when the underlying field is a Galois field $GF(2^n)$.

Security of ECC

- Just as RSA depends on the difficulty of large-number factorization for its security, ECC depends on the difficulty of the large number discrete logarithm calculation. This is referred to as the **Elliptic Curve Discrete Logarithm Problem** (ECDLP).

- It was shown by Menezes, Okamoto, and Vanstone (MOV) in 1993 that (for supersingular elliptic curves) the problem of solving the ECDLP problem (where the domain is the group $E_q(a, b)$) can be reduced to the much easier problem of finding logarithms in a finite field. There has been much work recently on extending the MOV reduction to general elliptic curves.

- In order to not fall prey to the MOV attack, the underlying elliptic curve and the base point chosen must satisfy what is known as the **MOV Condition**.

- The MOV condition is stated in terms of the **order** of the base point G. The order m of the base point G is the value of m such that $mG = \mathbf{O}$ where $\mathbf{O}$ is the additive identity element of the group $E_q(a, b)$.

- The MOV condition states that the **order** m of the base-point

should not divide $q^B - 1$ for small B, say for $B < 20$. Note that q is the prime p when the underlying finite field is Z_p or it is 2^n when the underlying finite field is $GF(2^n)$.

- When using $GF(2^n)$ finite fields, another security consideration relates to what is known as the **Weil descent attack**. To not be vulnerable to this attack, n must be a prime.

- Elliptic curves for which the total number of points on the curve equals the number of elements in the underlying finite field are also considered cryptographically weak.

Acknowledgements

I'd like to thank Helena Verrill and Subhash Kak for sharing their insights with me on the mathematics of elliptic curves and on the subject of elliptic curve cryptography. Helena Verrill is the source of much of the information provided regarding the singularity and supersingularity of elliptic curves. The curves shown on Slides 7 and 12 were generated by Jagadeesh Dyaberi using Matlab.