An Introduction to Elliptic Curves

Dr. Reza Akhtar

Algebra Short Course (SUMSRI)
Miami University
Summer 2002

Supplementary References

Silverman, Joseph H. and Tate, John. *Rational Points on Elliptic Curves.*
Silverman, Joseph H. *The Arithmetic of Elliptic Curves.*

The first book is a down-to-earth introduction to the study of elliptic curves with a focus on the arithmetic aspects of the theory (the interaction of number theory and geometry). It is aimed at the "math major" undergraduate audience and should be quite accessible. The second book is a graduate text; it contains most of the material in the first book but treats it in considerably greater generality, using more advanced algebraic techniques. Both of these books are exceptionally well written. These notes will follow the first source fairly closely.

1 Introduction

Number theory, broadly defined, is the study of the integers, which we usually represent by $\mathbf{Z}$. While the set of integers may seem innocuous at first, a bit of thought reveals that the structure of this set is mysterious and quite complex. Take, for example, the subset $P \subseteq \mathbf{Z}$ of prime numbers. Our knowledge of P is very limited; we know, for example, that there are infinitely many prime numbers – here is a nice proof due to Euclid:

Suppose, towards a contradiction, that there are only finitely many prime numbers, call them $p_1, p_2, \ldots, p_n$. Now consider the number

$$N = p_1 p_2 \ldots p_n + 1$$

This number N is obviously larger than any of the primes p_i, but it can't be prime since we're assuming that the p_i are the only primes around. Therefore, N is composite, and as such must be divisible by some prime number p_i. This means that $N = m \cdot p_i$, where m is some integer. Substituting this expression for N into the equation above, we get

$$m \cdot p_i = p_1 p_2 \dots p_n + 1$$

or

$$p_i(m - p_1 \dots p_{i-1} p_{i+1} \dots p_n) = 1$$

This is bad news: there's no way that a prime number (which has to be greater than 1) could divide 1! So we've reached a contradiction and there have to be infinitely many prime numbers.

Now that we've established that there are infinitely many prime numbers, there are deeper questions that beg to be answered. For example, 11 and 13 form a pair of prime numbers which differ by two (such pairs are called *twin primes*): are there infinitely many such pairs? Goldbach has conjectured that there are infinitely many such pairs, but the bottom line is that we still don't have a proof of this fact. Moreover, the techniques involved in making whatever progress has been achieved towards the solution of this problem come from all areas of mathematics: algebra, analysis, topology, combinatorics, to name a few.

Exercise. A *prime triple* is a sequence p_1, p_2, p_3 of prime numbers such that $p_3 = p_2 + 2$ and $p_2 = p_1 + 2$. Prove that $3, 5, 7$ is the only prime triple. (Hint: consider divisibility by 3.)

A very active area of research within number theory is that of *Diophantine Equations*, the study of integer or rational solutions of polynomial equations. For example, one could fix an integer $n \geq 3$ and ask which integers X, Y, and Z satisfy the equation

$$X^n + Y^n = Z^n$$

Equivalently, we could divide everything by Z^n and define new variables $x = X/Z$ and $y = Y/Z$ to rewrite the equation as

$$(X/Z)^n + (Y/Z)^n = 1 \text{ or } x^n + y^n = 1$$

(So the hunt for integer solutions (X, Y, Z) of the original equation is equivalently to that for rational solutions (x, y) of the latter equation)
As it turns out, the only solutions occur when one of X, Y and Z is zero; this is *Fermat's Last Theorem*, a result conjectured by Fermat in 1689 and proven by Andrew Wiles in 1993. Note that if we modify the problem and consider the case $n = 2$, the situation changes completely: now we're searching for so-called *Pythagorean triples* – $(3, 4, 5)$ and $(5, 12, 13)$ are among the (infinitely many) solutions to this equation.

Exercise.
Show that the equations $x^2 + y^2 = -7$ and $x^2 + y^2 = 100000000002$ have no solutions in integers x, y.

Here's another example: fix a rational number c and consider *Bachet's equation*:

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

We could ask: which *rational* numbers x and y satisfy this equation? This isn't so easy to solve directly, but we can make some interesting observations. Suppose that we know that (x, y) is a solution to the above equation. It isn't hard to verify by direct computation that

$$(\frac{x^4 - 8cx}{4y^2}, \frac{-x^6 - 20cx^3 + 8c^2}{8y^3})$$

is also a solution of the Bachet equation! (This is often referred to as the *duplication formula*. Better yet, since x and y are rational numbers, the coordinates of the above expression will still be rational numbers. So we can start with one rational solution (x_0, y_0), substitute x_0 for x and y_0 for y into the above formula to get a new rational solution (x_1, y_1), then substitute x_1 for x and y_1 for y to get yet another rational solution (x_2, y_2) to the Bachet equation, and so on. In fact, if $c \neq 1, -432$, it turns out that one can generate infinitely many rational solutions in this manner.
As an example, let's consider $c = -2$, i.e. the Bachet equation which reads

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

It's easy to see that $(3, 5)$ is a solution. Substituting into the duplication formula, we see that

$$(\frac{129}{100}, \frac{-383}{1000}) \text{ and } (\frac{2340922881}{7660^2}, \frac{113259286337292}{7660^3})$$

are also solutions. It's clear that the numbers get complicated very quickly!

We could also ask: which *integer* solutions x and y satisfy the Bachet equation? This is actually a much harder problem; Axel Thue showed in 1908 that there are only *finitely many* integer solutions.

Let's go back to the search for rational solutions. How on earth did Bachet come up with the duplication formula? The answer is, from the geometry! Suppose $P = (x_0, y_0)$ is a solution to $y^2 - x^3 = c$ which is a *rational point*, i.e. a point on the curve $y^2 = x^3 + c$ (pictured below) with x_0 and y_0 rational numbers: one could draw the tangent line through P and look at the point $Q = (x_1, y_1)$ at which it intersects the curve again. Since a line should, in principle, intersect a cubic in three points, and a tangency is counted as a "double intersection", such a point Q should exist (although it might be P again!).

Let's carry out these computations explicitly:
the tangent line L through P has equation $y = mx + b$, where m is the slope. To find the slope to the curve, we use implicit differentiation on the Bachet equation to get $2y\frac{dy}{dx} = 3x^2$ or $\frac{dy}{dx} = \frac{3x^2}{2y}$. So the slope m at P is $\frac{3x_0^2}{2y_0}$, and $b = y_0 - mx_0 = y_0 - \frac{3x_0^3}{2y_0} = \frac{2y_0^2 - 3x_0^3}{2y_0}$. We still need to find the coordinates of Q: to do this, we intersect the line $y = mx + b$ with the curve $y^2 = x^3 + c$; substituting for y, we get:

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

or

$$x^3 - m^2x^2 - 2bmx + (b^2 - c) = 0$$

What are the solutions for x here? Well, we know two of them, since the line intersects the curve "twice" (tangency) at P. So two of the solutions for x

are x_0. Now observe that the sum of the roots of the equation is the negative of the coefficient of x^2, so $x_0 + x_0 + x_1 = m^2$, or

$$x_1 = m^2 - 2x_0 = \frac{9x_0^4}{4y_0^2} - 2x_0 = \frac{9x_0^4 - 8x_0y_0^2}{4y_0^2}$$

From the Bachet equation, $x_0^3 = y_0^2 - c$, or $x_0^4 = xy_0^2 - cx_0$, so the above expression becomes

$$= \frac{8x_0^4 + x_0^4 - 8x_0y_0^2}{4y_0^2} = \frac{8x_0y_0^2 - 8cx_0 + x_0^4 - 8x_0y_0^2}{4y_0^2} = \frac{x_0^4 - 8cx_0}{4y_0^2}$$

which is exactly the first coordinate of the duplication formula. To get the second coordinate, simply observe that $y_1 = mx_1 + b$, throw our expressions for m, b, and x_1 into this equation, work through all the algebra, and you'll get the second coordinate of the duplication formula.

Exercise.
Make these vague remarks precise; that is, prove that

$$y_1 = \frac{-x^6 - 20cx^3 + 8c^2}{8y^3}.$$

Thus, the moral of the story is that given a solution to the Bachet equation (i.e. a point on the associated curve with rational coordinates), one can generate another solution by means of the simple geometric procedure described above.

Our short-term goal is to show that we can do something even more impressive in a more general setting; namely, given any plane curve $f(x, y) = 0$ defined by a cubic equation, we can make the rational points on the curve into a *group*.

2 Some Group Theory

2.1 Definitions

Let S be a set.

Definition 2.1. *A* binary operation *on S is a map $* : S \times S \longrightarrow S$. In plain language, a binary operation is a rule which inputs two elements a, $b \in S$ and outputs an element $a * b \in S$.*

For example, we may endow the set of natural numbers $\mathbf{N} = \{0, 1, 2, 3, \ldots\}$ with the binary operation $+$ (addition), or with the binary operation $\cdot$ (multiplication). However, $-$ (subtraction) is not a binary operation on $\mathbf{N}$, because $1 - 2 = -1$ is not in $\mathbf{N}$. Nevertheless, $-$ is a binary operation on $\mathbf{Z}$.

Definition 2.2. *A binary operation $*$ on a set S is called* associative *if for every $a, b, c \in S$, we have $(a * b) * c = a * (b * c)$.*

Clearly, addition (or multiplication) defines an associative binary operation on $\mathbf{N}$; subtraction, however, does *not* define an associative binary operation on $\mathbf{Z}$.

Definition 2.3. *Let S be a set with a binary operation $*$. An element $e \in S$ is called an* identity element *if for every $a \in S$, we have $e * a = a$ and $a * e = a$.*

In the above examples, 0 is an identity element (in fact, the only one) for $\mathbf{N}$ under addition, and 1 is the unique identity element for $\mathbf{N}$ under multiplication. As another example, consider the set of 2 by 2 matrices with real number entries. Matrix multiplication is a binary operation on this set, and the matrix

is an identity element. One could also consider the same set with componentwise addition as the binary operation; in that case, the identity element would be:

Definition 2.4. *Let S be a set with a binary operation $*$ and an identity element e. Let $a \in S$ be an element. An element $b \in S$ is called an* inverse *for a if $a * b = e$ and $b * a = e$.*

Consider the set $\mathbf{R}^*$ of all nonzero real numbers. Multiplication defines a binary operation on this set; 1 is the (only) identity element and the inverse of $a \in \mathbf{R}^*$ is $\dfrac{1}{a}$. (this makes sense since $a \neq 0$) As another example, one could consider the integers $\mathbf{Z}$ under addition. In this case, 0 is the only identity element and the inverse of $a \in \mathbf{Z}$ is $-a$.

Definition 2.5. *A* group *is a nonempty set S together with a binary operation $*$ satisfying:*

- $*$ *is associative*
- *There exists an identity element* $e \in S$ *for* $*$.
- *Every element of* S *has an inverse with respect to* $*$.

If we dispense with the third condition and only require the first two, we get what is called a *monoid*. If we dispense with the latter two conditions and only require the first, we get a *semigroup*. If we dispense with all three, we get a *magma*. (I have no idea where this name comes from)

Examples.

- $\mathbf{Z}$ is a group with respect to addition, but not with respect to multiplication (only 1 and -1 have inverses)
- Fix an integer n. The set of *residue classes mod* n, written $\mathbf{Z}_n$ is defined by $\{\bar{0}, \bar{1}, \ldots, n \bar{-} 1\}$. We can define a binary operation on $\mathbf{Z}_n$ by stipulating adding classes "modulo n". For example, if $n = 5$, then $\mathbf{Z}_n = \{\bar{0}, \bar{1}, \ldots, \bar{4}\}$ and we can define $\bar{1} + \bar{2} = \bar{3}$, $\bar{3} + \bar{4} = \bar{2}$ (because $3 + 4 = 7$, which is the same as 2 modulo 5).

  It's clear from this definition that addition is associative, that $\bar{0}$ is the (only) identity element and that the inverse of $\bar{a}$ is $n \bar{-} a$.
- The set $GL_2(\mathbf{R})$ of invertible 2 by 2 matrices forms a group under matrix multiplication. Indeed, if A and B are invertible matrices, then AB is also invertible; its inverse is $B^{-1}A^{-1}$, so matrix multiplication defines a binary operation. Once again, the matrix I_2 is the (only) identity element and all the inverse of an element $A \in GL_2(\mathbf{R})$ is just the inverse matrix A^{-1}.

Definition 2.6. *A group* G *is called* abelian *or* commutative *if for every* $a, b \in G$, *we have* $a * b = b * a$.

For example, $\mathbf{Z}$ with respect to addition is abelian; however, $GL_2(\mathbf{R})$ is not abelian. Notice how the definition of group is formulated to *exclude* the commutative condition so as to accomodate examples such as $GL_2(\mathbf{R})$.

Remark.
For groups G, we often write ab instead of $a * b$ for the binary operation. If the group G is abelian, we usually write $a + b$ instead of $a * b$, and label the identity element as 0 instead of e. Also, we write a^2 for $a * a$, a^{-2} for $a^{-1} * a^{-1}$, etc.

2.2 Elementary Properties

Here are some nice properties of groups. The first eliminates the need to keep on saying "the only identity element":

Proposition 2.7. *In a group G, the identity element is unique.*

Proof.
Suppose e and f are both identity elements for G.
Viewing f as an identity element and e as an arbitrary element, we have $e = ef$. Viewing e as an identity element and f as an arbitrary element, we have $ef = f$. Putting these two equations together yields $e = f$.

Now we check that inverses, too, are unique.

Proposition 2.8. *In a group G, each element has a unique inverse.*

Proof.
Fix an element $g \in G$ and suppose that $h \in G$ and $h' \in G$ are both inverses of g. Then $gh = hg = e$ and $gh' = h'g = e$.
Thus $h = he = h(gh') = (hg)h' = eh' = h'$.
(Note how we used the associative property in the third equality)

Definition 2.9. *Let G be a group and $a \in G$ an element. The* order *of a, written $|a|$, is the smallest positive integer m (if it exists) such that $a^m = e$. If no such integer exists, we say that a has infinite order.*

For example, the order of the identity element of any group is always 1; in fact, it is the only element of order 1. In $\mathbf{Z}$, every nonzero element has infinite order. (Right?) In $\mathbf{Z}_n$, every element has order $\leq n$; in fact, every element has order dividing n.
Exercise.
Show that in a group G, $|a| = |a^{-1}|$.

Definition 2.10. *Let G be a finite group. The* cardinality *or* order *of G, denoted $|G|$, is the number of elements in G.*

The following proposition establishes a link between the order of a group and the order of elements within that group.

Proposition 2.11. *Let G be a finite group and $a \in G$ an element. Then $|a|$ divides $|G|$.*

Definition 2.12. *Let G be a group. A* subgroup *of G is a subset $H \subseteq G$ which is a group in its own right.*

The next proposition follows easily from the definitions; we omit the proof.

Proposition 2.13. *Let G be a group. A subset $H \subseteq G$ is a subgroup if and only if H is nonempty, closed under the group operation, and closed under inverses.*

Exercise.
(a) Let G be an *abelian group* and $n \geq 1$ an integer. Prove that the set

$$G_n = \{x \in G : |x| \text{ divides } n\}$$

is a subgroup of G.
(b) Find an example of a nonabelian group G and an integer n such that G_n is not a subgroup of G.

Definition 2.14. *A group G is said to be* generated *by a subset $S \subseteq G$ if every element of G can be obtained by composing some (finite) combination of elements of S and their inverses together.*

For example, the group $\mathbf{Z}$ is generated by $\{1\}$, since every positive integer can be obtained by adding 1 to itself finitely many times; 0 is expressible as $1 + (-1)$ and every negative integer can be obtained by adding -1 to itself finitely many times. A group which has a generating subset consisting of one element is called a *cyclic* group. A group which has a finite generating subset is called a *finitely generated* group.

As another example, consider the group $\mathbf{Z} \times \mathbf{Z}$; this is the set of pairs (m, n) where $m, n \in \mathbf{Z}$; group composition is defined by coordinatewise addition, i.e. $(m, n) + (p, q) = (m + p, n + q)$. It's clear that this composition law is associative, that $(0, 0)$ is the identity element, and that $(-a, -b)$ is the inverse of (a, b). However, a bit of thought reveals that this group is not cyclic. (Why?) On the other hand, it is finitely generated, a set of generators being $\{(1, 0), (0, 1)\}$.
Exercise.
Find another set of generators for $\mathbf{Z} \times \mathbf{Z}$.

Definition 2.15. *Let G and G' be groups. A* homomorphism *is a map $\phi : G \longrightarrow G'$ such that for all $x, y \in G$,*

$$\phi(xy) = \phi(x)\phi(y)$$

A silly-looking but important example of a homomorphism is the *identity map* $id_G : G \longrightarrow G$ from a group to itself; this is defined by $id_G(g) = g$ for all $g \in G$.

For example, if V and W are (real) vector spaces, the conditions in the definition of a vector space imply that V is a group with respect to $+$ (addition of vectors) and similarly for W. Recall that a *linear transformation* is a map $T : V \longrightarrow W$ such that $T(a + b) = T(a) + T(b)$ and $T(ca) = cT(a)$ for all $a, b \in V$ and $c \in \mathbf{R}$. The first condition shows that T is actually a homomorphism.

To give another example, consider the "doubling map" $D : \mathbf{Z} \longrightarrow \mathbf{Z}$ given by the formula $D(n) = 2n$. Then, for any $a, b \in \mathbf{Z}$, we have

$$D(a + b) = 2(a + b) = 2a + 2b = D(a) + D(b)$$

and so D is a homomorphism.

Definition 2.16. *Let G and G' be groups. A homomorphism $\phi : G \longrightarrow G'$ is called an* isomorphism *if there exists a homomorphism $\psi : G' \longrightarrow G$ such that $\phi \circ \psi = id_{G'}$ and $\psi \circ \phi = id_G$. In this case we say that G and G' are* isomorphic *as groups.*

Groups which are isomorphic are considered to be essentially the same. For example, consider the set of rotations of the plane $G_4 = \{i, \rho_{\pi/2}, \rho_\pi, \rho_{3\pi/2}\}$, where i is the map which leaves everything alone and ρ_θ is counterclockwise rotation of the plane around the origin through an angle of θ radians. The set G_4 forms a group under composition of transformations; i is clearly the identity, and $\rho_{2\pi-\theta}$ is the inverse of ρ_θ. Furthermore, the group G_4 is isomorphic to $\mathbf{Z}_4$. One such isomorphism is obtained by sending i to $\bar{0}$, $\rho_{\pi/2}$ to $\bar{1}$, ρ_π to $\bar{2}$ and $\rho_{3\pi/2}$ to $\bar{3}$. The inverse map is defined by switching these definitions, and also defines a homomorphism, as you may check.

Exercise.
Suppose G, H, and J are groups. Let $\phi : G \longrightarrow H$ and $\psi : H \longrightarrow J$ be homomorphisms. Prove that the composition $\psi \circ \phi : G \longrightarrow J$ is also a homomorphism.

Exercise.
(a) Let G be a group and $a, b \in G$ elements. Prove that

$$(ab)^{-1} = b^{-1}a^{-1}.$$

(b) Suppose G is a group. Prove that the function

$$\iota : G \longrightarrow G$$

defined by $\iota(g) = g^{-1}$ is a homomorphism if and only if G is an abelian group.

3 Elliptic Curves and the Group Law

3.1 Projective Space

The Euclidean plane, $\mathbf{R}^2$, is often referred to as "2-dimensional affine space over the reals"; another common notation used for this set is $\mathbf{A}^2_{\mathbf{R}}$. Precisely,

$$\mathbf{A}^2_{\mathbf{R}} = \{(x, y) : x, y \in \mathbf{R}\}$$

This notation allows some flexibility; we could replace $\mathbf{R}$ by $\mathbf{C}$ and define

$$\mathbf{A}^2_{\mathbf{C}} = \{(x, y) : x, y \in \mathbf{C}\}$$

We could do the same thing with $\mathbf{Q}$, $\mathbf{Z}$, etc.

We're going to define a slightly more complicated variant of this object.
Consider the set $L = \{(X, Y, Z) \mid X, Y, Z \in \mathbf{R} \text{ and } X, Y, Z \text{ not all zero }\}$.
(This set is essentially $\mathbf{A}^3_{\mathbf{R}} - \{(0, 0, 0)\}$)
Define an equivalence relation on L by stipulating that $(A, B, C) \sim (\lambda A, \lambda B, \lambda C)$ for all nonzero real numbers lambda. For instance, $(\frac{1}{4}, \frac{1}{2}, \frac{2}{3})$ is equivalent to $(3, 6, 8)$ (use $\lambda = 12$).
The set $L/\sim$ of equivalence classes with respect to $\sim$ is called 2-*dimensional projective space over* $\mathbf{R}$ and is denoted $\mathbf{P}^2_{\mathbf{R}}$. This object is described as 2-dimensional because we started L, an essentially 3-dimensional object, and considered equivalence classes with respect to a linear (1-dimensional) relation. The equivalence class of (A, B, C) in $\mathbf{P}^2_{\mathbf{R}}$ is typically written $[A : B : C]$ to avoid confusion with affine space.

In summary, we have

$$\mathbf{P}^2_{\mathbf{R}} = \{[X : Y : Z] \mid X, Y, Z \in \mathbf{R} \text{ and } X, Y, Z \text{ not all zero } \}$$

We can partition the elements of $\mathbf{P}^2_{\mathbf{R}}$ into two subsets; the first, C_1 consists of those elements $[X : Y : Z]$ such that $Z \neq 0$ and the second, C_0, consists of those elements of the form $[X : Y : 0]$. It's clear from the definition of $\sim$ that no element of C_1 is equivalent to any element of C_0.

Let's examine elements of C_1 more closely. Since $Z \neq 0$, a typical element $[X : Y : Z]$ of C_1 is equivalent (take $\lambda = \frac{1}{Z}$) to $[\frac{X}{Z} : \frac{Y}{Z} : 1]$. So the elements of C_1 are essentially of the form $[x : y : 1]$, where x and y are allowed to be any real numbers! So, C_1 is somewhat like a copy of $\mathbf{A}^2_{\mathbf{R}}$ sitting inside $\mathbf{P}^2_{\mathbf{R}}$.

Now let's look at elements of C_0. The typical element of C_0 looks like $[X : Y : 0]$. Note that either X or Y is nonzero (remember, we excluded the triple consisting of all zeros) – if $X \neq 0$, then $[X : Y : 0]$ is equivalent to $[1 : \frac{Y}{X} : 0]$, which is essentially a copy of $\mathbf{A}^1_{\mathbf{R}}$. If $X = 0$, then the typical point has the form $[0 : Y : 0]$, which is equivalent to $[0 : 1 : 0]$ since Y is nonzero.
Thus, C_0 is a union of $\mathbf{A}^1_{\mathbf{R}}$ and the point $[0 : 1 : 0]$; by the above reasoning, this is itself a copy of $\mathbf{P}^1_{\mathbf{R}}$.

In conclusion: the set $\mathbf{P}^2_{\mathbf{R}}$ is a disjoint union of $\mathbf{A}^2_{\mathbf{R}}$ and $\mathbf{P}^1_{\mathbf{R}}$; the former consists of all (equivalence classes of) points of the form $[X : Y : Z]$ with $Z \neq 0$ and the latter consists of (equivalence classes of) points of the form $[X : Y : 0]$. The latter set is often referred to as the "line at infinity", since it's one-dimensional and isn't part of $\mathbf{A}^2_{\mathbf{R}}$.

3.2 Weierstrass Equations

Let's consider the set of points

$$E = \{[X : Y : Z] \in \mathbf{P}^2_{\mathbf{R}} : Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3\}$$

where a, b, c are real numbers.
We know that $\mathbf{P}^2_{\mathbf{R}}$ is a disjoint union of $\mathbf{A}^2_{\mathbf{R}}$ and the line at infinity; let's study the intersection of E with each such piece. When does E intersect the line at infinity? Well, precisely when the condition $Z = 0$ holds, in addition to the equation which defines E. Substituting, we get $0 = X^3$, which forces

$X = 0$; Y is allowed to be anything. Thus, E intersects the line at infinity in the points $[0 : Y : 0]$; however, since $Y \neq 0$, these are all (by the equivalence relation) the same point $[0 : 1 : 0]$. So, E intersects the line at infinity in the single point $\mathcal{O} = [0 : 1 : 0]$.

Now, to study the intersection of E with $\mathbf{A}^2_{\mathbf{R}}$ we need to see what happens when $Z \neq 0$. Well, then $[X : Y : Z]$ is the same point as $[\frac{X}{Z} : \frac{Y}{Z} : 1]$. Let $x = \frac{X}{Z}$ and $y = \frac{Y}{Z}$. Dividing the original equation for E by Z^3, we get

$$\frac{Y^2}{Z^2} = \frac{X^3}{Z^3} + a\frac{X^2}{Z^2} + b\frac{X}{Z} + c$$

or

$$y^2 = x^3 + ax^2 + bx + c$$

So, points on the intersection of E with $\mathbf{A}^2_{\mathbf{R}}$ are exactly solutions of $y^2 = x^3 + ax^2 + bx + c$. This latter equation is called the *dehomogenization* of the equation of E.

The moral: we can consider the curve E as the set of solutions to the so-called *Weierstrass equation* $y^2 = x^3 + ax^2 + bx + c$, together with a "point at infinity" $\mathcal{O}$, which sits outside the affine plane.

3.3 The Group Law

In this section, we'll show that (under certains conditions) we can make the points on E into a group $E(\mathbf{R})$, and that the points with *rational* coordinates form a subgroup $E(\mathbf{Q})$ of this group.

Definition 3.1. *Let E be a cubic curve with Weierstrass equation $y^2 = f(x) = x^3 + ax^2 + bx + c$. Let $F(x, y) = y^2 - f(x)$. The curve E is called* nonsingular *if there is no point on E at which the partial derivatives*

$$\frac{\partial F}{\partial x} = -f'(x) \text{ and } \frac{\partial F}{\partial y} = 2y$$

vanish simultaneously.
A nonsingular curve of this form is called an elliptic curve.

What is the significance of this condition? Well, starting from the equation $y^2 = f(x)$, we can try to find a formula for the slope $\dfrac{dy}{dx}$ of the tangent line to E at (x, y). Using implicit differentiation, we get $2y\dfrac{dy}{dx} = f'(x)$

$$\frac{dy}{dx} = \frac{f'(x)}{2y} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}}$$

This expression certainly makes sense if the denominator is nonzero, and can be interpreted as the "slope" of a vertical line if the numerator is nonzero and the denominator zero. However, if both quantities vanish, there really isn't a well-defined slope at that point.

In terms of the algebra, though, what happens if both partial derivatives vanish at some point $P = (x_0, y_0)$? Well, we have $f'(x_0) = 0$ and $2y_0 = 0$; so in particular, $y_0 = 0$. However, since $y^2 = f(x)$, it follows that $f(x_0) = 0$. So we have both $f(x_0) = 0$ and $f'(x) = 0$. By a result which we'll prove later (Proposition 3.2), it follows that $f(x)$ has a repeated root α. The converse is also true: if $f(x)$ has a repeated root x_0, then the partial derivatives of F will both vanish at the corresponding point $P = (x_0, y_0)$ on E.

Exercise.
Prove that the conic given by the equation

$$x^2 - 3xy + 2y^2 - x + 1 = 0$$

is nonsingular.

From now on, we're going to assume that E is a nonsingular curve with Weierstrass equation in the above form. As discussed above, this means that at least one of the partial derivatives does not vanish at each point of E; alternately, the right hand side $f(x)$ of the Weierstrass equation for E has distinct roots.

Let's start with two points $P, Q \in E$. For convenience, let's suppose for the time being that $P \neq Q$ and that neither point is equal to $\mathcal{O}$, the point at infinity. How could we "add" these points to get a new point? Let's take the line through P and Q and see where it intersects E; call this third point of intersection $P * Q$. One could ask: does the operation $*$ make the points of E into a group? Unfortunately, the answer is no; a little thought reveals

that there is no identity element. However, we can remedy the situation relatively easily: define $P + Q$ to be the reflection of $P * Q$ in the x-axis (i.e. if $P * Q = (x, y)$, then $P + Q = (x, -y)$). Of course, we need to cover other cases, too: if $P = Q$, then we should use the tangent line to P in place of the line between P and Q. Furthermore, if one of P and Q is $\mathcal{O}$, we stipulate $P + \mathcal{O} = P$ and $\mathcal{O} + Q = Q$. We'll see that $+$ makes E into a group.

Using this definition, it's easy to see that $+$ defines a binary operation on E and that $\mathcal{O}$ is the identity element. What is the inverse of a point $P = (x, y)$? Well, from the geometry one can see that $-P = (x, -y)$, is also a point of E, and that $P + (-P) = \mathcal{O} = (-P) + P$. The only snag is the associativity of $+$. Using the explicit formulas we're about to develop, one can check that $+$ is associative; however, it will probably take you several days to consider all of the side cases. Using more advanced techniques from algebraic geometry, it is possible to verify associativity of $+$ with considerably less pain, so we'll be satisfied by asserting that the operation is truly associative, but that the proof is truly painful.

Just as we did for the Bachet curve, we may also develop explicit formulas for the group law on E. Here we'll sketch how to do it for distinct points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$, neither of which is the point $\mathcal{O}$ at infinity. If we label $P * Q = (x_3, y_3)$, then $P + Q = (x_3, -y_3)$ by our definition. Now, P_3 is the third point of intersection of the line connecting P and Q to the curve E. We may calculate the slope of this line by using the slope formula

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1}$$

and the equation of the line $y = \lambda x + \nu$ using point-slope form. Now, we substitute this expression for y into the Weierstrass equation for E to get:

$$(\lambda x + \nu)^2 = x^3 + ax^2 + bx + c$$

or

$$x^3 + (a - \lambda^2)x^2 + (b - 2\lambda\nu)x + (c - \nu^2) = 0$$

As before, we note that x_3 is, along with x_1 and x_2 a root of this equation, and since the sum of the roots is $\lambda^2 - a$, this allows us to calculate x_3. We may then find y_3 by substituting x_3 for x in the equation of the line connecting P and Q.

Exercise.
Find explicit formulas for x_3 and y_3 in terms of $x_1, x_2, y_1, y_2, a, b, c$.

Somehow, the point is that the expressions that you will come up with for x_3 and y_3 are *rational* functions of $x_1, x_2, y_1,$ and y_2. In particular, if these latter four quantities are rational (i.e. P and Q are rational points), then x_3 and y_3 will be rational; thus, $P + Q$ is also a rational point. This shows that the set of rational points is closed under the group composition law. (Well, technically, we still need to consider side cases like $P = Q$, $P = \mathcal{O}$, etc, but the proof proceeds similarly in those cases) In fact, in the case $P = Q = (x, y)$ the formula for the x-coordinate of $P + Q = 2P$ is

$$x(2P) = \frac{x^4 - 2bx^2 - 8cx + b^2 - 4ac}{4x^3 + 4ax^2 + 4bx + 4c}$$

It's clear, too, that if $P = (x, y)$ is a rational point, then $-P = (x, -y)$ must also be a rational point. By Proposition 2.13, the subset

$$E(\mathbf{Q}) = \{(x, y) : \ y^2 = x^3 + ax^2 + bx + c : \ x, y \text{ rational } \} \cup \{\mathcal{O}\}$$

forms a subgroup of E.

Finally, it should be mentioned that a group law can be defined on any curve defined by a cubic equation in x and y. However, using a series of coordinate changes, one may always reduce the study of rational points on an arbitrary cubic to the study of rational points on a cubic defined by a Weierstrass equation. (This is a not-terribly-deep-but-somewhat-messy result from algebraic geometry) Thus, we will concentrate our attention on cubics defined by Weierstrass equations.

Exercise.
Verify the "duplication" formula asserted above.

Exercise.
Consider the point $P = (3, 8)$ on the cubic curve $y^2 = x^3 - 43x + 166$. Compute P, $2P$, $3P$, $4P$, and $8P$. Comparing $8P$ with P, what can you conclude?

3.4 Points of Order Two and Three

A natural question which arises in this context is the following. Consider a nonsingular cubic curve E given by a Weierstrass equation $y^2 = f(x) =$

$x^3 + ax^2 + bx + c$. Which points on E have order two? Alternatively, find all points $P = (x, y)$ such that $P \neq \mathcal{O}$ but $2P = \mathcal{O}$. By adding $-P$ to both sides of the above equation, we may write $P = -P$. Using our formula for the inverse of P, this now reads $(x, y) = (x, -y)$; thus $y = -y$ and so $y = 0$. How many values of x are there corresponding to $y = 0$? Well, look at the Weierstrass equation

$$0 = y^2 = x^3 + ax^2 + bx + c$$

It all depends on how many real roots the cubic on the right has. Certainly every cubic has at least one real root, and the other two roots are either two real numbers or complex conjugates. So, in terms of real coordinates, $E(\mathbf{R})$ has either one or three points of order 2; if we allow complex coordinates, $E(\mathbf{C})$ has three points of order 2. In any case, allowing complex coordinates, E has *four* points of order *dividing* 4.

As an example, let's look at E given by the Weierstrass equation $y^2 = x^3 + 8$. (Don't forget the point at infinity, though!) The cubic on the right, $x^3 + 8 = (x + 2)(x^2 - 2x + 4)$, has only one real root $x = -2$; the other two roots are given by the quadratic formula: $x = 1 \pm \sqrt{-3}$. So $E(\mathbf{R})$ has one point of order two and $E(\mathbf{C})$ has three points of order 2.

On the other hand, if we look at E given by $y^2 = (x-1)(x-2)(x-3)$, then the right hand side has three real roots, so in this case $E(\mathbf{R})$ has 3 points of order 2.

Now let's look at points of order 3. This isn't so bad, either, because the relation $3P = \mathcal{O}$ is equivalent to $2P = -P$. In particular, this means that $x(2P)$ and $x(-P)$ are the same, but since $x(-P) = x(P)$, this just means that $x(2P) = x(P)$. Fortunately, we have a nice formula for $x(2P)$ given in the last section: if we abbreviate $x(P)$ by x, then

$$x(2P) = \frac{x^4 - 2bx^2 - 8cx + b^2 - 4ac}{4x^3 + 4ax^2 + 4bx + 4c}$$

Setting this equal to x and solving (doing out the algebra), we conclude that

$$\psi_3(x) = 3x^4 + 4ax^3 + 6bx^2 + 12cx + (4ac - b^2) = 0$$

Thus, $P = (x, y)$ is a point of order three if and only if x satisfies the above equation.

How many points $P = (x, y)$ of order three are there? Well, the polynomial $\psi_3(x)$ has at most four roots, so there are at most four possibilities for x. For each point $P = (x, y)$ of order 3, it isn't hard to check that $-P = (x, -y)$ also has order 3, so these points come in pairs. Now, P and $-P$ are distinct points as long as $y \neq 0$. However, we can safely assume $y \neq 0$, because all points with $y = 0$ have order 2, as we saw above.

So the number of points of order three is equal to twice the number of *distinct* roots of $\psi_3(x)$.

Proposition 3.2. *The polynomial $\psi_3(x)$ has four distinct roots.*

To prove this statement, we need the following useful lemma:

Lemma 3.3. *A polynomial $f(x)$ has a repeated root α if and only if α is a common root of $f(x)$ and $f'(x)$ (its formal derivative)*

Proof.
Suppose $f(x)$ and $f'(x)$ share a common root, say α. Then $(x - \alpha)$ divides $f(x)$, so we can write

$$f(x) = (x - \alpha)g(x)$$

Using the product rule, we have

$$f'(x) = g(x) + (x - \alpha)g'(x)$$

Since $(x - \alpha)$ also divides $f'(x)$, this means that

$$(x - \alpha)h(x) = f'(x) = g(x) + (x - \alpha)g'(x)$$

Rearranging terms, we get

$$g(x) = (x - \alpha)(h(x) - g'(x))$$

In other words, $(x-\alpha)$ divides $g(x)$. Going back to the original factorization of $f(x)$, this means that $(x - \alpha)^2$ divides $f(x)$ and α is a repeated root of $f(x)$; so the roots of $f(x)$ are not distinct.

Conversely, suppose some root α of $f(x)$ occurs with multiplicity ≥ 2. In other words,

$$f(x) = (x - \alpha)^2 g(x)$$

Using the product rule again (and the chain rule!), we have

$$f'(x) = 2(x-\alpha)g(x) + (x-\alpha)^2 g'(x) = (x-\alpha)(2g(x) + (x-\alpha)g'(x))$$

and so α is also a root of $f'(x)$.

Corollary 3.4. *If $f(x)$ and $f'(x)$ share no common roots, then $f(x)$ has distinct roots.*

We now return to the proof of Proposition 3.2.

Recall how we obtained the polynomial $\psi_3(x)$. We set $x(2P) = x(P)$. However, we also have:

$$x(2P) = \frac{f'(x)^2}{4f(x)} - a - 2x$$

so

$$\psi_3(x) = 2f(x)f''(x) - f'(x)^2$$

(You could also just check this latter equation directly)
To check that $\psi_3(x)$ has distinct roots, it suffices by Corollary 3.4 to check that $\psi_3(x)$ and $\psi_3'(x)$ have no common roots. However,

$$\psi_3'(x) = 2f(x)f'''(x) = 12f(x)$$

So it a common root of $\psi_3(x)$ and $\psi_3'(x)$ would also be a comon root of $\psi(x) = 2(x)f''(x) - f'(x)$ and $12f(x)$, which in turn would be a common root of $f(x)$ and $f'(x)$. However, this is not possible, because it would contradict our assumption of the nonsingularity of E.

We may summarize our findings as follows:

Proposition 3.5. *Let E be an elliptic curve. Then E has exactly 8 points (allowing complex coordinates) of order 3 and 9 points of order dividing 3.*

Notice that, allowing complex coordinates, there is 1 point of order dividing 1, 4 points of order dividing 2, 9 points of order dividing 3. The following theorem (whose proof is beyond the scope of this discussion) should come as no surprise:

Theorem 3.6. *Let E be an elliptic curve and n a positive integer. Then the number of points on E of order dividing n is equal to n^2. In fact, this set of points forms a subgroup of E isomorphic to $\mathbf{Z}_n \times \mathbf{Z}_n$.*

Exercise.
Let E be an elliptic curve given by the usual Weierstrass equation

$$y^2 = f(x) = x^3 + ax^2 + bx + c.$$

(a) Prove that

$$\frac{d^2y}{dx^2} = \frac{2f''(x)f(x) - f'(x)^2}{4yf(x)} = \frac{\psi_3(x)}{4yf(x)}.$$

(b) Use this to deduce that a point $P = (x, y) \in E$ is a point of order three if and only if $P \neq \mathcal{O}$ and P is a point of inflection on the curve E.

(c) Now suppose $a, b, c \in \mathbf{R}$. Prove that $\psi_3(x)$ has exactly two real roots, say α_1, α_2 with $\alpha_1 < \alpha_2$. Prove that $f(\alpha_1) < 0$ and $f(\alpha_2) > 0$. Use this to deduce that the points in $E(\mathbf{R})$ of order dividing 3 form a cyclic group of order three.

4 The Nagell-Lutz Theorem

Let E be an elliptic curve with Weierstrass equation $y^2 = f(x) = x^3 + ax^2 + bx + c$; assume further that a, b, c are all *rational* numbers. Write $a = a_1/a_2$, $b = b_1/b_2$, $c = c_1/c_2$ where the numerator and denominator of each expression is an integer, and let d be large integer which is a multiple of each of the denominators a_2, b_2, and c_2. (For example, $d = a_2b_2c_2$ will do) Making a change of variables, we set $X = d^2x$ and $Y = d^3y$. Then our equation becomes:

$$Y^2 = X^3 + d^2aX^2 + d^4bX + d^6c$$

In particular, each of the coefficients is an *integer*. So we'll assume from now on that our Weierstrass equation has integer coefficients.

Our aim is to prove a theorem, due to Nagell and Lutz, that gives us a recipe for finding *all* the rational points of finite order. In particular, it also tells us that there are only finitely many such points.

4.1 The Discriminant

The *discriminant* of $f(x)$ is the quantity

$$D = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$$

(Note that if $a = 0$, this reduces to $-4b^3 - 27c^2$ – you may have seen this before in the context of cubic equations)
If we factor $f(x)$ over the complex numbers:

$$f(x) = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)$$

it isn't hard to check (by brute force computation) that

$$D = (\alpha_1 - \alpha_2)^2(\alpha_1 - \alpha_3)^2(\alpha_2 - \alpha_3)^2$$

Since we're assuming that E is nonsingular, i.e. that the roots of $f(x)$ are distinct, this implies in particular that D is nonzero.
Once again, direct computation shows that we have the following explicit formula:

$$D = ((18b-6a^2)x-(4a^3-15ab+27c))f(x)+((2a^2-6b)x^2+(2a^3-7ab+9c)x+(a^2b+3ac-4b^2)f'(x))$$

Collapsing some of these ugly expressions into more compact notation, we observe that there exist polynomials $r(x)$ and $s(x)$ with *integer* coefficients such that

$$D = r(x)f(x) + s(x)f'(x)$$

This formula is useful in proving the following surprising lemma:

Lemma 4.1. *Let $P = (x, y)$ be a point on E such that both P and $2P$ have integer coordinates. Then either $y = 0$ or $y|D$.*

Proof.
We start by assuming that $y \neq 0$ and prove that $y|D$. Because $y \neq 0$, we know that P does not have order 2 and therefore that $2P \neq \mathcal{O}$, so we may write $2P = (X, Y)$. By assumption, x, y, X, Y are all integers. The duplication formula asserts that

$$2x + X = \lambda^2 - a, \text{ where } \lambda = \frac{f'(x)}{2y}$$

Since x, X and a are all integers, it follows that λ^2 and hence λ is an integer. Finally, since $2y$ and $f'(x)$ are integers, too, we see that $2y|f'(x)$. In particular, $y|f'(x)$. However, $y^2 = f(x)$, so $y|f(x)$, too. Now we use the relation

$$D = r(x)f(x) + s(x)f'(x)$$

The coefficients of r and s are integers, so in particular $r(x)$ and $s(x)$ are both integers. Finally, since y divides both $f(x)$ and $f'(x)$, this formula shows that y divides D, too, as deisred.

At last we are ready to state the Nagell-Lutz Theorem:

Theorem 4.2. *(Nagell-Lutz)*
Let

$$y^2 = f(x) = x^3 + ax^2 + bx + c$$

be an elliptic curve with integer coefficients a, b, c; let D be the discriminant of the cubic polynomial. Let $P = (x, y)$ be a rational point of finite order. Then x and y are integers. If $y = 0$ then P has order 2; otherwise, y divides D.

Remark.
Note that the Nagell-Lutz Theorem is not an "if and only if" statement: just because a point may have integer coordinates does **not** mean that it has finite order. Furthermore, Lemma 4.1 shows that to prove the Nagell-Lutz theorem, it suffices to show that x and y are integers. The last statement will then follow from this lemma.

Now suppose we are trying to classify all the points $P = (x, y)$ of finite order on some elliptic curve E. Well, we already know about the points of order 2. The Nagell-Lutz Theorem tells us that we after computing the value of the discriminant D (from the Weierstrass equation), the only y-values we need consider are *integers* which divide D. There are obviously only finitely many such values, so we just need to comb through the list and see which ones give rise to rational points lying on E.
Remark.
The Nagell-Lutz Theorem may be strengthened to stipulate that y^2 has to divide D. This form is more useful for computational purposes.
Example.

Find all points of finite order on the curve $y^2 = x^3 + 4x$.

First, it's clear that the point $\mathcal{O}$ at infinity, which has projective coordinates $[0 : 1 : 0]$, is a rational point. We continue by classifying the points of order 2. Recall that these are determined by setting $x^3 + 4x = 0$, i.e. $x(x^2 + 4) = 0$. This yields one point: $(0, 0)$ (remember that $(2i, 0)$, $(-2i, 0)$, where $i = \sqrt{-1}$ are not rational points).

To proceed further, we compute the discriminant D: in our case, $D = -256$. Now we can use the (strong form of the) Nagell-Lutz Theorem; any rational point $P = (x, y)$ which does not have order 1 or 2 must have x and y both integers and y^2 dividing -256. Fortunately, $256 = 2^8$, so this means

$$y = \pm 1, \pm 2, \pm 4, \pm 8, \pm 16$$

Note that we only have to worry about testing the positive values for y, since $P = (x, y)$ is a rational point if and only if $-P = (x, -y)$ is a rational point.

Let's work through these cases separately. If $y = 1$, then we're trying solving the equation $1 = x^3 + 4x$ for *integers* x. Factoring, we get $1 = x(x^2 + 4)$, and it's clear that no integer satisfies this equation. If $y = 2$, we get $4 = x(x^2 + 4)$; there aren't any solutions to this equation, either. If $y = 4$, we get $16 = x(x^2 + 4)$. Since x has to be an integer, x has to divide 16, so are possibilities are $x = \pm 1, \pm 2, \pm 4, \pm 8, \pm 16$. We can eliminate the negative values instantly, since they would make the right hand side of the equation negative. Combing through the possibilites, we find that $x = 2$ is one such value satisfies this equation! So we've found a point $(2, 4)$, along with its inverse $(2, -4)$. Since there are no other values of x satisfying this equation, we move on. If $y = 8$, we get $64 = x(x^2 + 4)$. Again, we can argue that x must divide 64 and that x must be positive. By repeated trial, we find that none of these values satisfy this equation. Finally, if $y = 16$, we get $256 = x(x^2 + 4)$. Once again, we find that x must be positive and divide 256, and that no such values satisfy this equation.

Thus, the rational points on E are:

$$E(\mathbf{Q}) = \{\mathcal{O}, (0, 0), (2, 4), (2, -4)\}$$

The question is, what are the orders of these elements? We know $\mathcal{O}$ has order 1 and $(0, 0)$ has order 2, and that the other two elements have order greater than 2. Since $E(\mathbf{Q})$ is a group of order 4, Proposition 2.11 tells us that $(2, 4)$

and $(2, -4)$ have order dividing 4; since we know already that they each have order greater than 2, this forces (each of) them to have order 4. We conclude that this group is in fact isomorphic to $\mathbf{Z}_4$.

What remains of the proof of the Nagell-Lutz theorem – namely, that the coordinates (x, y) of a rational point on an elliptic curve are in fact *integers*, is a straightforward, but computationally intricate argument. Unfortunately, it requires some knowledge of quotient groups and special properties of homomorphisms which we can't get into here for want of time. So, we'll just leave the Nagell-Lutz Theorem for now and move on.

Before leaving the Nagell-Lutz Theorem, we note that this theorem gives a bound (namely the discriminant D) on the y-coordinate of a rational point of finite order on the curve E, but that this bound depends on E. (Remember, we calculated D as a function of the coefficients in the Weierstrass equation for E) One could also ask: exactly which orders are possible? This question was studied for a long time; the answer was finally provided by Barry Mazur in the following (very difficult) theorem:

Theorem 4.3. *(Mazur) Let E be an elliptic curve defined over $\mathbf{Q}$. (i.e. the coefficients in the Weierstrass equation for E may be chosen to lie in $\mathbf{Q}$) Suppose that $E(\mathbf{Q})$ contains a point of finite order m. Then*

$$1 \leq m \leq 10 \quad or \quad m = 12$$

Exercise.
Verify the formula given above for the discriminant.

Exercise.
For each of the following curves, determine all of the points of finite order, and determine the order of each such point. Feel free to use the strong form of the Nagell-Lutz Theorem.
(a) $y^2 = x^3 - 2$. (b) $y^2 = x^3 + 1$. (c) $y^2 - y = x^3 - x^2$.

5 Mordell's Theorem

The result we are after is the following:

Theorem 5.1. *Let E be an elliptic curve whose Weierstrass equation has rational coefficients. Then the group $E(\mathbf{Q})$ is finitely generated.*

This is truly an amazing theorem. It states that given just finitely many (rational) points on E, one can get *any* rational point on E by adding or subtracting some combination of these (finitely many) points.

As in the case of the Nagell-Lutz Theorem, we will not prove the Mordell Theorem *per se*, but rather give some indication of how it proceeds. First, however, we give some background:

5.1 Cosets: more abstract algebra

Definition 5.2. *Let G be a subgroup and $H \subseteq G$ a subgroup. Let $a \in G$ be any element. The set*

$$aH = \{ah : h \in H\}$$

is called a coset *of H in G. The element a is called a* representative *for this coset.*

To fix some ideas, here's an example. Let $G = \mathbf{Z}$ and $H = 2\mathbf{Z} = \{2n : n \in \mathbf{Z}\} = \{\text{all even integers}\}$. Then H satisfies the requirements for being a subgroup of G. We could consider the coset $0 + H$ (we're writing $0 + H$ instead of $0H$ since we're using the addition symbol '+' instead of multiplication to represent the group law). As a set $0 + H = \{0 + h : h \in H\}$ is simply H, the even integers. On the other hand, $1 + H = \{1 + h : h \in H\}$ is the set of all odd integers. What is $2 + H$? Well, if you think about it for a minute, it'll become clear that $2 + H = H$, also. More generally, $n + H$ is the set of even numbers if n is even; it is the set of odd numbers if n is odd.

Notice how there are only two distinct cosets of H in G. We make the following definition:

Definition 5.3. *Let G be a group and $H \subseteq G$ a subgroup. The* index *of H in G, written $[G : H]$ is the number of distinct cosets of H in G, if this quantity is finite. If not, we say that H has infinite index in G.*

Cosets have some nice properties – we won't need the following proposition in our later discussions, but it is interesting in its own right and casts some light on the general situation.

Proposition 5.4. *Let H be a subgroup of a group G and $a, b \in G$ two elements. Then:*

1. *Either* $aH = bH$ *or* $aH \cap bH = \emptyset$.

2. $aH = H$ *if and only if* $a \in H$.

3. $aH = bH$ *if and only if* $b^{-1}a \in H$.

Proof.
To prove the first assertion, suppose $aH \cap bH \neq \emptyset$, so let $x \in aH \cap bH$. This means that x may be written $x = ah$ for some $h \in H$ and as $x = bh'$ for some $h' \in H$. Thus $ah = bh'$, and multiplying both expressions on the right by h^{-1}, we have $ahh^{-1} = bh'h^{-1}$ or $a = bh'h^{-1}$. Since H is a subgroup and $h \in H$, it follows that $h^{-1} \in H$, too. Furthermore, since $h' \in H$ and $h^{-1} \in H$, the fact that H is a subgroup means that $h'h^{-1}$ is also a member of H. Thus, $a \in bH$ and hence $aH \subseteq bH$. By symmetric reasoning, $bH \subseteq aH$. Thus, $aH = bH$.

To prove the second assertion, first suppose that $aH = H$. Since $e \in H$, it follows that $a = ae \in aH$, which by hypothesis is just H. Thus $a \in H$. Conversely, suppose that $a \in H$. Then, since H is a subgroup, it's clear that $aH \subseteq H$. Furthermore, the fact that H is a subgroup implies that $a^{-1} \in H$, so any element $h \in H$ may be written as $h = a(a^{-1}h)$. Since the element in parentheses is an element of H, it follows that $h \in aH$, and hence that $H \subseteq aH$.

The third assertion follows almost immediately from the second:

$$aH = bH \text{ iff } b^{-1}aH = b^{-1}bH \text{ iff } b^{-1}aH = H$$

By the second proposition this is true if and only if $b^{-1}a \in H$.

Corollary 5.5. *With notation as above, the cosets of* H *in* G *partition* G*; that is, every element of* G *is contained in exactly one coset of* H.

Proof.
Indeed, $a \in G$ is contained in aH, and by the first assertion of Proposition 5.4, this is the only one.

5.2 Heights and Descent

Heights are devices for measuring the arithmetic "size" of a rational number. What do we mean by "arithmetic size"? Well, the numbers 1 and 57/58 are almost equal in absolute value, but in terms of prime factorizations, $57/58 = 3 \cdot 19/2 \cdot 29$ is much more complicated than 1. So, we want the height to be some measure of the complexity of the fractional representation of a rational number; to this end we define, for a rational number $x = \frac{m}{n}$ written in lowest terms,

$$H(x) = H(\frac{m}{n}) = \max\{|m|, |n|\}$$

In our example above, $H(1) = 1$ and $H(57/58) = 58$. One of the most important properties of the height, however, is the following proposition, whose proof happens to be self-evident:

Proposition 5.6. *Let $n \in Z$ be a positive integer. Then $\{x \in \mathbf{Q} : H(x) \leq n\}$ is a finite set.*

Exercise.
Prove that the set of rational numbers x with height $H(x) \leq k$ contains at most $2k^2 + k$ elements.

We can play the same game with elliptic curves. Given an elliptic curve E and a rational point $P = (x, y) \in E(\mathbf{Q})$, we can simply define

$$H(P) = H(x).$$

We will see later that the height behaves somewhat multiplicatively in the sense that $H(P + Q)$ is comparable (but not quite equal) to $H(P)H(Q)$. Sometimes it's more convenient to work with a function that behaves additively; hence we define

$$h(P) = \log H(P).$$

This function $h(P)$ is called the *logarithmic height*; from basic properties of logarithms, it follows that $h(P + Q) = \log H(P + Q)$ is comparable to $\log H(P)H(Q) = h(P) + h(Q)$. Note that h always takes on nonnegative real values, since $H(P)$ can never be less 1, regardless of the value of P. In any case, notice that for any real number $M > 0$, we still have the property that

$$\{P \in E(\mathbf{Q}) : H(P) \leq M\}$$

is a finite set, and similarly if we replace $H(P)$ with $h(P)$. This is because $H(P) \leq M$ means that the x-coordinate of P must be a fraction m/n with $|m|, |n| \leq M$ – there are only finitely many such x, and to each such x there are at most two corresponding y-values. The definition as it stands takes care of all the points on E except the point at infinity; for this, we simply declare by fiat that

$$H(\mathcal{O}) = 1; \quad \text{equivalently,} \quad h(\mathcal{O}) = 0$$

From here, the proof of the Mordell-Weil Theorem follows from four lemmas, together with the descent theorem. We list the lemmas below; following a brief discussion of them, we state and explain the significance of the descent theorem.

In all the statements below, E is an elliptic curve with Weierstrass equation $y^2 = x^3 + ax^2 + bx + c$.

Lemma 5.7. *For every real number M, the set*

$$\{P \in E(\mathbf{Q}) : h(P) \leq M\}$$

is finite.

Lemma 5.8. *Let P_0 be a (fixed) rational point on E. There is a constant k_0, depending on P_0 and a, b, c, so that:*

$$h(P + P_0) \leq 2h(P) + k_0 \text{ for all } P \in E(\mathbf{Q})$$

Lemma 5.9. *There is a constant k, depending on a, b, c, so that:*

$$h(2P) \geq 4h(P) - k \text{ for all } P \in E(\mathbf{Q})$$

Lemma 5.10. *The index $[E(\mathbf{Q}) : 2E(\mathbf{Q})]$ is finite. (Here we're using the notation $2E(\mathbf{Q})$ to denote the set of points of $E(\mathbf{Q})$ which are twice other points).*

Note that we have already proved Lemma 5.7 in our discussion above. Lemma 5.8 basically says that adding a fixed point P_0 to any point P essentially doubles the value of h, although we're still forced to correct the difference with the (essentially harmless) constant k_0. Lemma 5.9, on the other hand, says

that *doubling* a point increases its (logarithmic) height by a factor of (essentially) 4; in other words, the height acts somewhat like a quadratic form. Lemma 5.10 is really the deepest (and hardest) result here; it is sometimes referred to as the "weak Mordell theorem", since it is in some sense the most crucial ingredient in the proof of the Mordell Theorem.

The engine that drives the proof of Mordell's Theorem is the following "descent theorem". The name "descent theorem" was chosen because the proof is very much in the spirit of Fermat's method of infinite descent. Roughly speaking, one starts with an arbitrary point and tries to produce an infinite sequence of successively smaller points (size being measured by the height function); eventually, one is led to one of two conclusions: either the group is finitely generated or one reaches a contradiction in producing smaller points, since the height, being an integer, cannot be less than 1.

Without further ado, here is the statement of the theorem, stated in somewhat more generality than we require. Mordell's Theorem will follow immediately from the descent theorem, applied to $\Gamma = E(\mathbf{Q})$, in view of the four preceding lemmas.

Theorem 5.11. *Let Γ be an abelian group. Suppose there is a function $h : \Gamma \longrightarrow [0, \infty)$ with the following three properties:*

- *For every real number M, the set $\{P \in \Gamma : h(P) \leq M\}$ is finite.*
- *For every $P_0 \in \Gamma$, there is constant k_0 so that*

$$h(P + P_0) \leq 2h(P) + k_0 \text{ for all } P \in \Gamma.$$

- *There is a constant k so that*

$$h(2P) \geq 4h(P) - k \text{ for all } P \in \Gamma.$$

Suppose further that the index $[\Gamma : 2\Gamma]$ is finite.

Then Γ is finitely generated.

Proof.
Since we know that $[\Gamma : 2\Gamma] = n$ is finite, choose representatives $Q_1, \ldots, Q_n$ for the (distinct) cosets of 2Γ in Γ. Now let $P \in \Gamma$ be an arbitrary element. Since $P + 2\Gamma$ is a coset, it must be one of $Q_1 + 2\Gamma, \ldots, Q_n + 2\Gamma$; say,

$$P + 2\Gamma = Q_{i_1} + 2\Gamma.$$

By the third assertion of Proposition 5.4, this is equivalent to asserting that:

$$P - Q_{i_1} \in 2\Gamma$$

or that there exists $P_1 \in \Gamma$ such that

$$P - Q_{i_1} = 2P_1$$

Now replace P with P_1 and keep repeating this procedure to get equations:

$$P_1 - Q_{i_2} = 2P_2$$
$$P_2 - Q_{i_3} = 2P_3$$

$$\ldots$$

$$P_{m-1} - Q_{i_m} = 2P_m$$

The essence of the proof is to show that since P_i is pretty much equal to $2P_{i+1}$, the height of P_{i+1} is (by the third assumption in the descent theorem) pretty much one fourth the height of P_i. Thus the sequence P, P_1, P_2 produces points of successively decreasing height, and eventually we'll end up with a set of points having bounded height, which by the first assumption is finite. We need to be more precise, though.

From the first equation, we have

$$P = Q_{i_1} + 2P_1.$$

Now substitute the second equation

$$P_1 = Q_{i_2} + 2P_2.$$

into this to get

$$P = Q_{i_1} + 2Q_{i_2} + 4P_2.$$

Continuing in a similar manner, we obtain

$$P = Q_{i_1} + 2Q_{i_2} + 4Q_{i_3} + \ldots + 2^{m-1}Q_{i_m} + 2^m P_m.$$

In particular, this shows that P is in the subgroup generated by P_m and the Q_i, $1 \leq i \leq n$.

Now let's apply the second assumption with $-Q_i$ playing the role of P_0; we then obtain a constant k_i such that

$$h(P - Q_i) \leq 2h(P) + k_i \text{ for all } P \in \Gamma.$$

If we do this for all $i = 1, \ldots, n$ and let $k_0 = \max\{k_1, \ldots, k_n\}$, we clearly have

$$h(P - Q_i) \leq 2h(P) + k_0 \text{ for all } P \in \Gamma \text{ and all } 1 \leq i \leq n.$$

Now apply the third assumption: we get a constant k such that

$$h(2P) \geq 4h(P) - k \text{ for all } P \in \Gamma.$$

Rearranging terms and using the specific point $P = P_j$, we have $4h(P_j) \leq h(2P_j) + k$. By the initial equations defining the P_j and Q_{i_j}, the expression on the right equals $h(P_{j-1} - Q_{i_j}) + k$, which by the second assumption is $\leq 2h(P_{j-1}) + k_0 + k$. Summarizing, we have a chain of inequalitites:

$$4h(P_j) \leq h(2P_j) + k = h(P_{j-1} - Q_{i_j}) + k \leq 2h(P_{j-1}) + k_0 + k.$$

Dividing by 4, we get

$$h(P_j) \leq \frac{1}{2}h(P_{j-1}) + \frac{k_0 + k}{4}$$

$$= \frac{3}{4}h(P_{j-1}) - \frac{1}{4}(h(P_{j-1}) - (k_0 + k)).$$

From this last equation, we see that if $h(P_{j-1}) \geq k_0 + k$, then

$$h(P_j) \leq \frac{3}{4}h(P_{j-1}).$$

We claim that there exists some m such that $h(P_m) \leq k_0 + k$. Let's examine the list:

$$P, P_1, P_2, \ldots$$

As long as P_{j-1} has height $\geq k_0 + k$, the above equation says that the next point will have height at most $\frac{3}{4}P_j$. However, repeatedly multiplying a number by $\frac{3}{4}$ will cause it to approach zero, so eventually we'll find an index m such that $h(P_m) \leq k_0 + k$.

Now we've shown that every element $P \in \Gamma$ may be written as

$$P = a_1 Q_1 + a_2 Q_2 + \ldots a_n Q_n + 2^m P_m$$

where $a_1, \ldots, a_n$ are integers (some of them could be zero), and P_m is a point satisfiying $h(P_m) \leq k_0 + k$. Hence the set

$$\{Q_1, \ldots, Q_n\} \cup \{R \in \Gamma : h(R) \leq k_0 + k\}$$

generates Γ. The first set is clearly finite and the second is finite by our first assumption. Thus, Γ is finitely generated.