Introductory Number Theory

Course No. 100 331

Spring 2006

Michael Stoll

Contents

1. Very Basic Remarks

The following properties of the integers $\mathbb{Z}$ are fundamental.

(1) $\mathbb{Z}$ is an integral domain (i.e., a commutative ring such that $ab = 0$ implies $a = 0$ or $b = 0$).
(2) $\mathbb{Z}_{\geq 0}$ is well-ordered: every nonempty set of nonnegative integers has a smallest element.
(3) $\mathbb{Z}$ satisfies the *Archimedean Principle*: if $n > 0$, then for every $m \in \mathbb{Z}$, there is $k \in \mathbb{Z}$ such that $kn > m$.

2. Divisibility

2.1. **Definition.** Let a, b be integers. We say that "a divides b", written

$$a \mid b\,,$$

if there is an integer c such that $b = ac$. In this case, we also say that "a is a divisor of b" or that "b is a multiple of a".

We have the following simple properties (for all $a, b, c \in \mathbb{Z}$).

(1) $a \mid a$, $1 \mid a$, $a \mid 0$.
(2) If $0 \mid a$, then $a = 0$.
(3) If $a \mid 1$, then $a = \pm 1$.
(4) If $a \mid b$ and $b \mid c$, then $a \mid c$.
(5) If $a \mid b$, then $a \mid bc$.
(6) If $a \mid b$ and $a \mid c$, then $a \mid b \pm c$.
(7) If $a \mid b$ and $|b| < |a|$, then $b = 0$.
(8) If $a \mid b$ and $b \mid a$, then $a = \pm b$.

2.2. **Definition.** We say that "d is the greatest common divisor of a and b", written

$$d = \gcd(a, b) \quad \text{or} \quad d = a \sqcap b\,,$$

if $d \mid a$ and $d \mid b$, $d \geq 0$, and for all integers k such that $k \mid a$ and $k \mid b$, we have $k \mid d$.

We say that "m is the least common multiple of a and b", written

$$m = \operatorname{lcm}(a, b) \quad \text{or} \quad m = a \sqcup b\,,$$

if $a \mid m$ and $b \mid m$, $m \geq 0$, and for all integers n such that $a \mid n$ and $b \mid n$, we have $m \mid n$.

In a similar way, we define the greatest common divisor and least common multiple for any set S of integers. We have the following simple properties.

(1) $\gcd(\emptyset) = 0$, $\operatorname{lcm}(\emptyset) = 1$.
(2) $\gcd(S_1 \cup S_2) = \gcd(\gcd(S_1), \gcd(S_2))$, $\operatorname{lcm}(S_1 \cup S_2) = \operatorname{lcm}(\operatorname{lcm}(S_1), \operatorname{lcm}(S_2))$.
(3) $\gcd(\{a\}) = \operatorname{lcm}(\{a\}) = |a|$.
(4) $\gcd(ac, bc) = |c| \gcd(a, b)$.
(5) $\gcd(a, b) = \gcd(a, ka + b)$.

3. The Euclidean Algorithm

How can we compute the gcd of two given integers? The key for this is the last property of the gcd listed above. In order to make use of it, we need the operation of division with remainder.

3.1. **Proposition.** *Given integers a and b with $b \neq 0$, there exist unique integers q ("quotient") and r ("remainder") such that $0 \leq r < |b|$ and $a = bq + r$.*

Proof. Existence: Consider $S = \{a - kb : k \in \mathbb{Z}, a - kb \geq 0\}$. Then $S \subset \mathbb{Z}_{\geq 0}$ is nonempty and therefore has a smallest element $r = a - qb$ for some $q \in \mathbb{Z}$. We have $r \geq 0$ by definition, and if $r \geq |b|$, then $r - |b|$ would also be in S, hence r would not have been the smallest element.

Uniqueness: Suppose $a = bq + r = bq' + r'$ with $0 \leq r, r' < |b|$. Then $b \mid r - r'$ and $0 \leq |r - r'| < |b|$, therefore $r = r'$. This implies $bq = bq'$, hence $q = q'$ (since $b \neq 0$). □

3.2. **Algorithm GCD (Euclidean Algorithm).** Given integers a and b, we do the following.

(1) Set $n = 0$, $a_0 = |a|$, $b_0 = |b|$.
(2) If $b_n = 0$, return a_n as the result.
(3) Write $a_n = b_n\, q_n + r_n$ with $0 \leq r_n < b_n$.
(4) Set $a_{n+1} = b_n$, $b_{n+1} = r_n$.
(5) Replace n by $n + 1$ and go to step 2.

We claim that the result returned is $\gcd(a, b)$. (Observe that $0 \leq b_{n+1} < b_n$ if the loop is continued, hence the algorithm must terminate.)

Proof. We show that for all n that occur in the loop, we have $\gcd(a_n, b_n) = \gcd(a, b)$. The claim follows, since the return value $a_n = \gcd(a_n, 0) = \gcd(a_n, b_n)$ for the last n. For $n = 0$, we have $\gcd(a_0, b_0) = \gcd(|a|, |b|) = \gcd(a, b)$. Now suppose that we know $\gcd(a_n, b_n) = \gcd(a, b)$ and that $b_n \neq 0$ (so the loop is not terminated). Then $\gcd(a_{n+1}, b_{n+1}) = \gcd(b_n, a_n - b_n q_n) = \gcd(b_n, a_n) = \gcd(a_n, b_n)$ (use property (5) of the gcd). □

3.3. **Theorem.** *Fix $a, b \in \mathbb{Z}$. The integers of the form $xa + yb$ with $x, y \in \mathbb{Z}$ are exactly the multiples of $d = \gcd(a, b)$. In particular, there are $x, y \in \mathbb{Z}$ such that $d = xa + yb$.*

Proof. Since d divides both a and b, d also has to divide $xa + yb$. So these numbers are multiples of d. For the converse, it suffices to show that d can be written as $xa + yb$. This follows by induction from the Euclidean Algorithm: Leet N be the last value of n. Then $d = a_N \cdot 1 + b_N \cdot 0$; and if $d = x_{n+1}a_{n+1} + y_{n+1}b_{n+1}$, then we have $d = y_{n+1}a_n + (x_{n+1} - q_n y_{n+1})b_n$, so setting $x_n = y_{n+1}$ and $y_n = x_{n+1} - q_n y_{n+1}$, we have $d = x_n a_n + y_n b_n$. So in the end, we must also have $d = x_0 a_0 + y_0 b_0$. □

There is a simple extension of the Euclidean Algorithm that also computes numbers x and y such that $\gcd(a, b) = xa + yb$. It looks like this.

3.4. **Algorithm XGCD (Extended Euclidean Algorithm).** Given integers a and b, we do the following.

(1) Set $n = 0$, $a_0 = |a|$, $b_0 = |b|$, $x_0 = \operatorname{sign}(a)$, $y_0 = 0$, $u_0 = 0$, $v_0 = \operatorname{sign}(b)$.
(2) If $b_n = 0$, return (a_n, x_n, y_n) as the result.
(3) Write $a_n = b_n\, q_n + r_n$ with $0 \leq r_n < b_n$.
(4) Set $a_{n+1} = b_n$, $b_{n+1} = r_n$, $x_{n+1} = u_n$, $y_{n+1} = v_n$, $u_{n+1} = x_n - u_n\, q_n$, $v_{n+1} = y_n - v_n\, q_n$.
(5) Replace n by $n + 1$ and go to step 2.

By induction, one shows that $a_n = x_n\, a + y_n\, b$ and $b_n = u_n\, a + v_n\, b$, so upon exit, the return values (d, x, y) satisfy $xa + yb = d = \gcd(a, b)$.

3.5. **Proposition.** *If n divides ab and $\gcd(n, a) = 1$, then n divides b.*

Proof. By Thm. 3.3, there are x and y such that $xa + yn = 1$. We multiply by b to obtain $b = xab + ynb$. Since n divides ab, n divides the right hand side and therefore also b. □

4. Prime Numbers and Unique Factorization

4.1. **Definition.** A positive integer p is called a "prime number" (or simply a "prime"), if $p > 1$ and the only positive divisors of p are 1 and p.

(This is really the definition of an "irreducible" element in a ring!)

4.2. **Proposition.** *If p is a prime and a, b are integers such that $p \mid ab$, then $p \mid a$ or $p \mid b$.*

(This is the general definition of a "prime" element in a ring!)

Proof. We can assume that $p \nmid a$ (otherwise we are done). Then $\gcd(a, p) = 1$ (since the only other positive divisor of p, namely p itself, does not divide a). Then by Prop. 3.5, p divides b. □

Now let $n > 0$ be a positive integer. Then either $n = 1$, or n is a prime number, or else n has a "proper" divisor d such that $1 < d < n$. Then we can write $n = de$ where also $1 < e < n$. Continuing this process with d and e, we finally obtain a representation of n as a product of primes. (If $n = 1$, it is given by an empty product (a product without factors) of primes.)

4.3. **Theorem.** *This prime factorization of n is unique (up to ordering of the factors).*

Proof. By induction on n. For $n = 1$, this is clear (there is only one empty set ...). So assume that $n > 1$ and that we have written

$$n = p_1 p_2 \dots p_k = q_1 q_2 \dots q_l$$

with prime numbers p_i, q_j, where $k, l \geq 1$. Now p_1 divides the product of the q_j's, hence p_1 has to divide one of the q_j's. Up to reordering, we can assume that $p_1 | q_1$. But the only positive divisors of q_1 are 1 and q_1, so we must in fact have $p_1 = q_1$. Let $n' = n/p_1 = n/q_1$, then

$$n' = p_2 p_3 \dots p_k = q_2 q_3 \dots q_l\,.$$

Since $n' < n$, we know by induction that (perhaps after reordering the q_j's) $k = l$ and $p_j = q_j$ for $j = 2, \dots, k$. This proves the claim. □

4.4. **Definition.** This shows that we can write every nonzero integer n uniquely as

$$n = \pm \prod_p p^{v_p(n)}$$

where the product is over all prime numbers p, and the exponents $v_p(n)$ are non-negative integers, all but finitely many of which are zero. $v_p(n)$ is called the "valuation of n at p".

We have the following simple properties.

(1) $v_p(mn) = v_p(m) + v_p(n)$.
(2) $m \mid n \iff \forall p : v_p(m) \le v_p(n)$.
(3) $\gcd(m,n) = \prod_p p^{\min(v_p(m),v_p(n))}$, $\quad \operatorname{lcm}(m,n) = \prod_p p^{\max(v_p(m),v_p(n))}$.
(4) $v_p(m+n) \ge \min(v_p(m), v_p(n))$, with equality if $v_p(m) \ne v_p(n)$.

Property (3) implies that $\gcd(m,n)\ \operatorname{lcm}(m,n) = mn$ for positive m, n. In general, we have $\gcd(m,n)\ \operatorname{lcm}(m,n) = |mn|$.

If we set $v_p(0) = +\infty$, then all the above properties hold for all integers (with the usual conventions like $\min\{e, \infty\} = e$, $e + \infty = \infty$, ...).

We can extend the valuation function from the integers to the rational numbers by setting

$$v_p\left(\frac{r}{s}\right) = v_p(r) - v_p(s)$$

(Exercise: check that this is well-defined). Properties (1) and (4) above then hold for rational numbers, and a rational number x is an integer if and only if $v_p(x) \ge 0$ for all primes p.

5. Congruences

5.1. **Definition.** Let a, b and n integers with $n > 0$. We say that "a is congruent to b modulo n", written

$$a \equiv b \bmod n\,,$$

if n divides the difference $a - b$.

5.2. **Congruence is an equivalence relation.**
For fixed n and arbitrary $a, b, c \in \mathbb{Z}$, we have:

(1) $a \equiv a \bmod n$.
(2) If $a \equiv b \bmod n$, then $b \equiv a \bmod n$.
(3) If $a \equiv b \bmod n$ and $b \equiv c \bmod n$, then $a \equiv c \bmod n$.

Hence we can partition $\mathbb{Z}$ into "congruence classes mod n": we let

$$\bar{a} = a + n\mathbb{Z} = \{a + nx : x \in \mathbb{Z}\} = \{b \in \mathbb{Z} : a \equiv b \bmod n\}$$

(in the $\bar{a}$ notation, n must be clear from the context) and

$$\mathbb{Z}/n\mathbb{Z} = \{\bar{a} : a \in \mathbb{Z}\}\,.$$

We then have

$$a \equiv b \bmod n \iff b \in \bar{a} \iff \bar{a} = \bar{b}\,.$$

5.3. **Proposition.** *The map*

$$\{0, 1, \ldots, n-1\} \longrightarrow \mathbb{Z}/n\mathbb{Z}\,, \quad r \longmapsto \bar{r} = r + n\mathbb{Z}$$

is a bijection. In particular, $\mathbb{Z}/n\mathbb{Z}$ has exactly n elements.

Proof. The map is clearly well-defined. It is injective: assume $\bar{r} = \bar{s}$ with $0 \le r, s < n$. Then $r \equiv s \bmod n$, so $n \mid r - s$ and $|r - s| < n$, therefore $r = s$. It is surjective: Let $\bar{a}$ be a congruence class and write $a = nq + r$ with $0 \le r < n$. Then $\bar{a} = \bar{r}$. □

Since the representative r of a class $\bar{a}$ is given by the (least nonnegative) residue of a when divided by n, congruence classes are also called "residue classes".

5.4. **The congruence classes form a commutative ring.**
We define addition and multiplication on $\mathbb{Z}/n\mathbb{Z}$:

$$\bar{a} + \bar{b} = \overline{a + b}\,, \qquad \bar{a} \cdot \bar{b} = \overline{ab}$$

We have to check that these operations are well-defined. This means that if $a \equiv a' \bmod n$ and $b \equiv b' \bmod n$, then we must have $a + b \equiv a' + b' \bmod n$ and $ab \equiv a'b' \bmod n$. Now

$$(a + b) - (a' + b') = (a - a') + (b - b') \quad \text{is divisible by } n,$$

and also

$$ab - a'b' = (a - a')b + a'(b - b') \quad \text{is divisible by } n.$$

Once these operations are well-defined, all the commutative ring axioms carry over immediately from $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$.

5.5. **Congruences are useful.** Why are congruences a useful concept? They give us a kind of "Mickey Mouse" image of the integers (lumping together many integers into one residue class, thus losing information), with the advantage that the resulting structure $\mathbb{Z}/n\mathbb{Z}$ has only finitely many elements. This means that all sorts of questions that are difficult to answer with respect to $\mathbb{Z}$ are effectively (though not necessarily efficiently, if n is large) decidable with respect to $\mathbb{Z}/n\mathbb{Z}$. If we can show in this way that something is impossible over $\mathbb{Z}/n\mathbb{Z}$, then this often implies a negative answer for $\mathbb{Z}$, too.

Consider, for example, the equation $x^2 + y^2 - 15\,z^2 = 7$. Does it have a solution in integers? That to decide seems to be hard. On the other hand, we can very easily make a table of all possible values of the left hand side in $\mathbb{Z}/8\mathbb{Z}$: it is easy to see that a square is always $\equiv 0$, 1, or 4 mod 8, and adding three of these values (note that $-15 \equiv 1 \bmod 8$) leads to all residue classes mod 8 with one exception — the left hand side is never $\equiv 7 \bmod 8$.

So a solution is not possible in $\mathbb{Z}/8\mathbb{Z}$. But any solution in $\mathbb{Z}$ would lead to an image solution in $\mathbb{Z}/8\mathbb{Z}$, hence there can be no solution in $\mathbb{Z}$ either.

6. Coprime Integers and Multiplicative Inverses

When does a class $\bar{a}$ have a multiplicative inverse in $\mathbb{Z}/n\mathbb{Z}$? We have to solve the congruence $ax \equiv 1 \bmod n$; equivalently, there need to exist integers x and y such that $ax + ny = 1$. By Thm. 3.3, this is equivalent with $\gcd(a, n) = 1$. In this case, we say that "a and n are relatively prime" or "a and n are coprime", and sometimes write $a \perp n$. We can use the extended Euclidean Algorithm to find the inverse.

6.1. **Theorem.** *The ring $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is a prime number.*

Proof. Clear for $n = 1$ (a field has at least two elements 0 and 1, and 1 is not a prime number). For $n > 1$, $\mathbb{Z}/n\mathbb{Z}$ is not a field if and only if there is some $a \in \mathbb{Z}$, not divisible by n, such that $d = \gcd(n, a) > 1$. This implies that $1 < d < n$ is a proper divisor of n, hence n is not a prime. Conversely, if d is a proper divisor of n, then $\gcd(n, d) = d$, and $\bar{d}$ is not invertible. □

If $\gcd(n, a) = 1$, then $\bar{a}$ is called a "primitive residue class $\bmod n$"; its uniquely determined multiplicative inverse in $\mathbb{Z}/n\mathbb{Z}$ is denoted $\bar{a}^{-1}$. The prime residue classes form a group, the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$ of the ring $\mathbb{Z}/n\mathbb{Z}$.

When $n = p$ is prime, then the field $\mathbb{Z}/p\mathbb{Z}$ is also denoted $\mathbb{F}_p$; we have $\mathbb{F}_p^\times = \mathbb{F}_p \setminus \{\bar{0}\}$ for the multiplicative group; in particular, $\#\mathbb{F}_p^\times = p - 1$.

6.2. **Definition.** The Euler ϕ function is defined for $n > 0$ by

$$\phi(n) = \#(\mathbb{Z}/n\mathbb{Z})^\times = \#\{a \in \mathbb{Z} : 0 \le a < n, a \perp n\}\,.$$

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\phi(n)$	1	1	2	2	4	2	6	4	6	4	10	4	12	6	8	8	16	6	18	8

We have that n is prime if and only if $\phi(n) = n - 1$.

6.3. **Proposition.** *If p is prime and $e \ge 1$, then $\phi(p^e) = p^{e-1}(p - 1)$.*

Proof. Clear for $e = 1$. For $e > 1$, observe that $a \perp p^e \iff a \perp p \iff p \nmid a$, so

$$\begin{aligned}\phi(p^e) &= \#\{a \in \mathbb{Z} : 0 \le a < p^e, p \nmid a\} \\ &= p^e - \#\{a \in \mathbb{Z} : 0 \le a < p^e, p \mid a\} \\ &= p^e - p^{e-1} = (p - 1)p^{e-1}\,.\end{aligned}$$

□

6.4. **A Recurrence.** Counting the numbers between 0 (inclusive) and n (exclusive) according to their gcd with n (which can be any (positive) divisor d of n), we obtain

$$\sum_{d|n} \phi\Big(\frac{n}{d}\Big) = \sum_{d|n} \phi(d) = n\,.$$

This can be read as a recurrence for $\phi(n)$:

$$\phi(n) = n - \sum_{d|n, d<n} \phi(d)\,.$$

We get

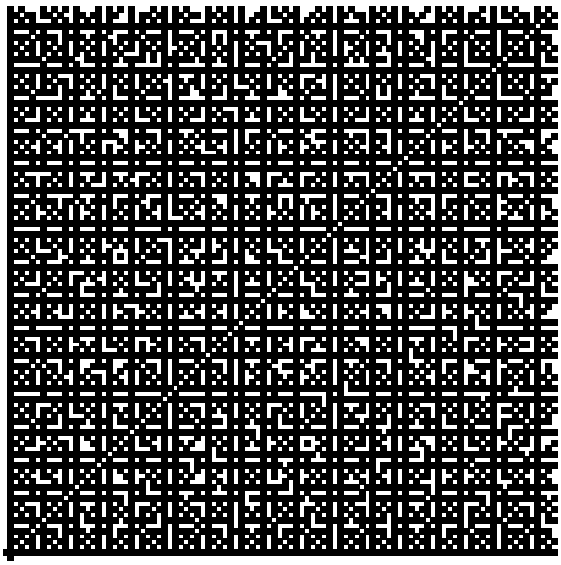
$$\phi(1) = 1 - 0 = 1\,, \quad \phi(2) = 2 - \phi(1) = 1\,, \quad \phi(3) = 3 - \phi(1) = 2\,,$$
$$\phi(4) = 4 - \phi(2) - \phi(1) = 2\,, \quad \phi(6) = 6 - \phi(3) - \phi(2) - \phi(1) = 2\,, \quad \text{etc.}$$

Obviously, the recurrence determines $\phi(n)$ uniquely: If integers a_n ($n \ge 1$) satisfy

$$\sum_{d|n} a_d = n\,,$$

then $a_n = \phi(n)$. This still holds if the n's are restricted to be divisors of some fixed integer N.

6.5. **A Question.** The following picture represents in black the coprime pairs (m, n) with $0 \leq m, n \leq 100$.



The black squares appear to be fairly evenly distributed, so the following question should make sense.

What is the probability that two random (positive) integers are coprime?

Take a large positive integer N and consider all pairs (m, n) with $1 \leq m, n \leq N$. Call $f(N)$ the number of such pairs with $m \perp n$. Since $m' \perp n'$ for all m, n, where $m' = m/\gcd(m, n)$, $n' = n/\gcd(m, n)$, we can count the pairs according to their gcd. We get

$$\begin{aligned} N^2 &= \#\{(m, n) : 1 \leq m, n \leq N\} \\ &= \sum_{g=1}^{N} \#\{(m, n) : 1 \leq m, n \leq N,\ \gcd(m, n) = g\} \\ &= \sum_{g=1}^{N} \#\{(m', n') : 1 \leq m', n' \leq N/g,\ m' \perp n'\} \\ &= \sum_{g=1}^{N} f\left(\lfloor N/g \rfloor\right) \end{aligned}$$

The probability we are looking for is $P = \lim_{N\to\infty} P(N)$, where $P(N) = f(N)/N^2$. We get

$$1 = \sum_{g=1}^{N} \frac{\lfloor N/g \rfloor^2}{N^2} P(\lfloor N/g \rfloor)\,.$$

Observe that the terms in the sum are between 0 and $1/g^2$, hence the sum is uniformly absolutely convergent. Passing to the limit as $N \to \infty$, we obtain

$$1 = \sum_{g=1}^{\infty} \frac{P}{g^2}\,,$$

hence

$$P = \frac{1}{\sum\limits_{g=1}^{\infty} g^{-2}} = \frac{6}{\pi^2} \approx 0.608\,.$$

(Exercise: where is the gap in this argument?)

We can apply what we have learned to *linear congruences.*

Given a, b, and n, what are the solutions x of

$$ax \equiv b \bmod n\,?$$

In other words, for which $x \in \mathbb{Z}$ does there exist a $y \in \mathbb{Z}$ such that $ax + ny = b$?

6.6. **Theorem.** *The congruence* $ax \equiv b \bmod n$ *has no solutions unless* $\gcd(a,n) \mid b$. *If there are solutions, they form a residue class modulo* $n/\gcd(a,n)$.

Proof. By Thm. 3.3, the condition $\gcd(a,n) \mid b$ is necessary and sufficient for solutions to exist. Let $g = \gcd(a,n)$. If $g \mid b$, we can divide the equation $ax+ny = b$ by g to get $a'x + n'y = b'$, where $a = a'g$, $n = n'g$, $b = b'g$. Then $\gcd(a',n') = 1$, and we can solve the equation $\bar{a}'\,\bar{x} = \bar{b}'$ for $\bar{x}$ (in $\mathbb{Z}/n'\mathbb{Z}$): $\bar{x} = \bar{b}'\,(\bar{a}')^{-1}$. So the set of solutions x is given by this residue class modulo n'. □

7. The Chinese Remainder Theorem

Now let us consider simultaneous congruences:

$$x \equiv a \bmod m\,, \qquad x \equiv b \bmod n$$

Is such a system solvable? What does the solution set look like? If $d = \gcd(m,n)$, then we obviously need to have $a \equiv b \bmod d$ for a solution to exist. On the other hand, when $m \perp n$, solutions do always exist.

7.1. **Chinese Remainder Theorem.** *If* $m \perp n$, *then the above system has solutions* x; *they form a residue class modulo* mn.

Proof. Since $m \perp n$, we can find u and v with $mu+nv = 1$ by Thm. 3.3. Consider $x = anv + bmu$. We have

$$x = anv + bmu \equiv anv = a - amu \equiv a \bmod m$$

and similarly $x \equiv b \bmod n$. So solutions exist. Now we show that y is anoher solution if and only if $mn \mid y - x$. If mn divides $y - x$, then m and n both divide $y - x$, hence $y \equiv x \equiv a \bmod m$ and $y \equiv x \equiv b \bmod n$. Now assume y is another solution. Then m and n both divide $y - x$. So $n \mid y - x = mt$. By Prop. 3.5, $n \mid t$ and hence $mn \mid mt = y - x$. □

There is a straight-forward extension to more than two simultaneous congruences.

7.2. **Chinese Remainder Theorem.** *If the numbers* m_1, m_2, …, m_k *are coprime in pairs, then the system of congruences*

$$x \equiv a_1 \bmod m_1\,, \quad x \equiv a_2 \bmod m_2\,, \ldots\,, \quad x \equiv a_k \bmod m_k$$

has solutions; they form a residue class modulo $m_1 m_2 \ldots m_k$.

Proof. Induction on k using Thm. 7.1. Note that $a \perp b$, $a \perp c$ implies $a \perp bc$. □

Now we can answer the question from the beginning of this section.

7.3. **Theorem.** *The system*

$$x \equiv a \bmod m\,, \qquad x \equiv b \bmod n$$

has solutions if and only if $a \equiv b \bmod \gcd(m,n)$*. If solutions exist, they form a residue class modulo* $\operatorname{lcm}(m,n)$*.*

Proof. We have already seen that the condition is necessary. Let $d = \gcd(m,n)$. We can find u and v such that $mu + nv = b - a$ by Thm. 3.3. Then $x = a + mu = b - nv$ is a solution. As in the proof of Thm. 7.1, we see that y is another solution if and only if m and n both divide $y - x$. This means that the solutions form a residue class modulo $\operatorname{lcm}(m,n)$. □

We can give the Chinese Remainder Theorem a more algebraic formulation.

7.4. **Theorem.** *Assume that* $m_1, m_2, \ldots, m_k$ *are coprime in pairs. Then the natural ring homomorphism*

$$\mathbb{Z}/m_1 m_2 \ldots m_k\mathbb{Z} \longrightarrow \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_k\mathbb{Z}$$

is an isomorphism. In particular, we have an isomorphism of multiplicative groups

$$(\mathbb{Z}/m_1 m_2 \ldots m_k\mathbb{Z})^\times \cong (\mathbb{Z}/m_1\mathbb{Z})^\times \times (\mathbb{Z}/m_2\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/m_k\mathbb{Z})^\times\,.$$

Proof. By the above, the homomorphism is bijective, hence an isomorphism. □

7.5. **A Formula for ϕ.** The Chinese Remainder Theorem 7.4 implies that

$$\phi(mn) = \phi(m)\,\phi(n) \qquad \text{if } m \perp n.$$

A similar formula holds for products with more factors. Applying this to the prime factorization of n, we get

$$\phi(n) = \prod_{p|n} p^{v_p(n)-1}(p-1) = n \prod_{p|n}\Big(1 - \frac{1}{p}\Big)\,.$$

8. Fermat's and Euler's Theorems

A very nice property of the finite fields $\mathbb{F}_p$ and all their extension fields is that the map $x \mapsto x^p$ is not only compatible with multiplication: $(xy)^p = x^p y^p$, but also with addition!

8.1. **Theorem ("Freshman's Dream").** *Let* F *be a field of prime characteristic* p *(this means that* $p \cdot 1_F = 0_F$*; for us, the basic example is* $F = \mathbb{F}_p$*). Then for all* $x, y \in F$*, we have* $(x+y)^p = x^p + y^p$*.*

Proof. By the Binomial Theorem,

$$(x+y)^p = x^p + \binom{p}{1}x^{p-1}y + \binom{p}{2}x^{p-2}y^2 + \cdots + \binom{p}{k}x^{p-k}y^k + \cdots + \binom{p}{p-1}xy^{p-1} + y^p.$$

Now the binomial coefficients $\binom{p}{k}$ for $1 \le k \le p-1$ all are integers divisible by p (why?), and since F is of characteristic p, all the corresponding terms in the formula vanish, leaving only $x^p + y^p$. □

We can use this to give one proof of the following fundamental fact.

8.2. **Theorem (Fermat's Little Theorem).** *Let p be a prime number. For all $\bar{a} \in \mathbb{F}_p$, we have $\bar{a}^p = \bar{a}$. (Equivalently, for all $a \in \mathbb{Z}$, p divides $a^p - a$.)*

Proof.
First Proof: By induction on a. $a = 0$ is clear. Now, by Thm. 8.1, p divides $(a+1)^p - a^p - 1$ for all $a \in \mathbb{Z}$. But then p also divides $((a+1)^p - (a+1)) - (a^p - a)$, hence:

$$p \mid a^p - a \iff p \mid (a+1)^p - (a+1)$$

This gives the inductive step upwards and downwards, hence the claim holds for all $a \in \mathbb{Z}$.

Second Proof: Easy proof using Algebra. $\bar{a} = \bar{0}$ is clear. Hence it suffices to show that $\bar{a}^{p-1} = \bar{1}$ for all $\bar{a} \neq \bar{0}$. This is a consequence of the fact that $\#\mathbb{F}_p^\times = p - 1$ and the general theorem that $g^{\#G} = 1$ for any g in any finite group G.

Third Proof: By Combinatorics (for $a > 0$). Consider putting beads that can have colors from a set of size a at p equidistant places around a circle (to form "necklaces"). There will be a^p necklaces in total, a of which will consist of beads of only one color. The remaining $a^p - a$ come in bunches of p, obtained by rotation, so p has to divide $a^p - a$. □

The algebra proof can readily be generalized.

8.3. **Theorem (Euler).** *Let n be a positive integer. Then for all $a \in \mathbb{Z}$ with $a \perp n$, we have $a^{\phi(n)} \equiv 1 \bmod n$.*

Proof. Under the assumption, $\bar{a} \in (\mathbb{Z}/n\mathbb{Z})^\times$, and by definition, $\#(\mathbb{Z}/n\mathbb{Z})^\times = \phi(n)$. By the general fact from algebra used in the second proof of Thm. 8.2, the claim follows. □

8.4. **Example.** What is $7^{11^{13}} \bmod 15$? By Thm. 8.3, $7^8 \equiv 1 \bmod 15$ (as $\phi(15) = 8$). On the other hand, $11 \equiv 3 \bmod 8$, and $3^4 \equiv 1 \bmod 8$ (in fact, already 3^2 is $\equiv 1 \bmod 8$). So $11^{13} = (11^4)^3 \cdot 11 \equiv 11 \equiv 3 \bmod 8$, and then $7^{11^{13}} \equiv 7^3 = 343 \equiv 13 \bmod 15$.

8.5. **A Consequence of Fermat's Little Theorem.** Consider the polynomial $X^p - X$ with coefficients in $\mathbb{F}_p$. By Fermat's Little Theorem 8.2, every element $\bar{a} \in \mathbb{F}_p$ is a root of this polynomial. Now $\mathbb{F}_p$ is a field, and so we can "divide out" the roots successively to find that

$$X^p - X = \prod_{\bar{a} \in \mathbb{F}_p} (X - \bar{a}) \,.$$

This implies that for any polynomial $f(X)$ dividing $X^p - X$ (in the polynomial ring $\mathbb{F}_p[X]$), the number of its distinct roots in $\mathbb{F}_p$ equals the degree $\deg f(X)$. More generally, if f is any polynomial in $\mathbb{F}_p[X]$, we can compute the number of distinct roots of f in $\mathbb{F}_p$ by the formula

$$\#\{\bar{a} \in \mathbb{F}_p : f(\bar{a}) = 0\} = \deg \gcd(f, X^p - X) \,.$$

9. Structure of $\mathbb{F}_p^\times$ and $(\mathbb{Z}/p^n\mathbb{Z})^\times$

Fermat's Theorem 8.2 tells us that the multiplicative order of any nonzero element $\bar{a}$ of $\mathbb{F}_p$ (this is the smallest positive integer n such that $\bar{a}^n = \bar{1}$) divides $p-1$. (The set of all n such that $\bar{a}^n = \bar{1}$ consists exactly of the multiples of the order.) Now the question arises, are there elements of order $p-1$? In other, more algebraic terms, is the group $\mathbb{F}_p^\times$ cyclic? The answer is "yes".

9.1. **Theorem.** The multiplicative group $\mathbb{F}_p^\times$ is cyclic. (In other words, there exist elements $\bar{g} \in F_p^\times$ such that all $\bar{a} \in \mathbb{F}_p^\times$ are powers of $\bar{g}$. The corresponding integers g are called "primitive roots mod p".)

Proof. Obviously, all elements of $\mathbb{F}_p^\times$ of order dividing d (where d is a divisor of $p-1$) will be roots of $X^d - \bar{1}$. Since d divides $p-1$, $X^d - \bar{1}$ divides $X^{p-1} - \bar{1}$ and hence also $X^p - X$ (as polyomials). By 8.5, it follows that $X^d - 1$ has exactly d roots in $\mathbb{F}_p$. Let a_d be the number of elements of exact order d. Then we get

$$d = \sum_{k|d} a_k \,.$$

By the statement in 6.4, it follows that $a_d = \phi(d)$; in particular, $a_{p-1} = \phi(p-1) \geq 1$. Hence primitive roots exist. □

9.2. **Examples.** The proof shows that there are exactly $\phi(p-1)$ essentially distinct primitive roots mod p. For the first few primes, we get the following table.

Prime	Primitive Roots
2	1
3	2
5	2, 3
7	3, 5
11	2, 3, 8, 9
13	2, 6, 7, 11

There is a famous conjecture, named after Artin, that asserts that every integer $g \neq -1$ that is not a square is a primitive root mod infinitely many different primes. (Why are squares no good?) This has been proven assuming another famous conjecture, the Extended Riemann Hypothesis. The best unconditional result so far seems to be that the statement is true for all allowed integers, with at most three exceptions. On the other hand, the statement is not known to hold for any particular integer g!

9.3. **Proposition.** Let G be a finite multiplicative abelian group of order n. An element $g \in G$ is a generator of G (and so G is cyclic) if and only if $g^{n/q} \neq 1_G$ for all prime divisors q of n.

Proof. If g is a generator, then n is the least positive integer m such that $g^m = 1_G$, hence the condition is necessary. Now if g is not a generator, then its order m divides n, but is smaller than n, hence m divides n/q for some prime divisor q of n. It follows that $g^{n/q} = 1_G$. □

Let us use this result to show that $(\mathbb{Z}/p^n\mathbb{Z})^\times$ is cylic if p is an odd prime and $n \geq 1$.

9.4. **Theorem.** Let g be a primitive root modulo p, where p is an odd prime. Then one of g and $g+p$ is a primitive root modulo p^n for all $n \geq 1$.

Proof. We know that $g^{p-1} = 1 + ap$ for some $a \in \mathbb{Z}$. If $p \nmid a$, let $h = g$; otherwise we set $h = g + p$; then we have $h^{p-1} = 1 + a'p$ with $p \nmid a'$:

$$(g+p)^{p-1} = g^{p-1} + (p-1)g^{p-2}p + bp^2 \equiv 1 - kp \bmod p^2$$

(with some integer b) where $k \equiv g^{p-2} \bmod p$ is not divisible by p. Hence we have in both cases that $h^{p-1} \equiv 1 + ap \bmod p^2$ with $p \nmid a$.

Now I claim that for all $n \geq 0$, we have

$$h^{p^n(p-1)} \equiv 1 + ap^{n+1} \bmod p^{n+2}\,.$$

This follows by induction from the case $n = 0$:

$$\begin{aligned} h^{p^{n+1}(p-1)} &= \left(h^{p^n(p-1)}\right)^p \\ &= (1 + (a+bp)p^{n+1})^p \\ &= 1 + p\,(a+bp)p^{n+1} + cp^{n+3} \\ &= 1 + ap^{n+2} + (b+c)p^{n+3} \end{aligned}$$

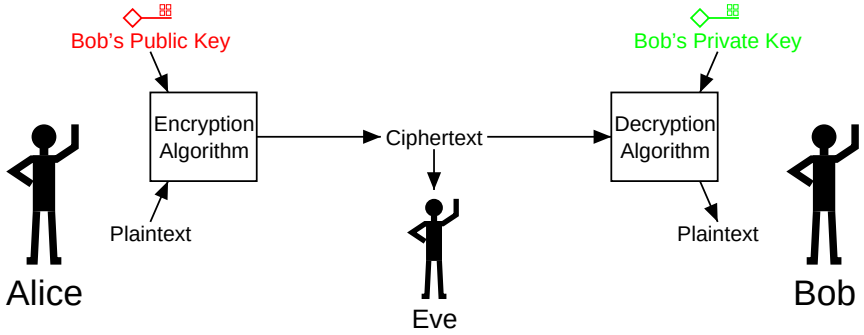
Here b and c are suitable integers, and the penultimate equality uses that $p \geq 3$ (since then the last term in the binomial expansion, $(a+bp)^p p^{(n+1)p}$, is divisible by p^{n+3}, as are the intermediate ones, even when $n = 0$).

Now let $n \geq 2$, and let q be a prime divisor of $\phi(p^n) = p^{n-1}(p-1)$. If q divides $p-1$, then $h^{(p-1)/q} \not\equiv 1 \bmod p$, hence also $h^{p^{n-1}(p-1)/q} \equiv h^{(p-1)/q} \not\equiv 1 \bmod p$, so $h^{p^{n-1}(p-1)/q} \not\equiv 1 \bmod p^n$. If $q = p$, then we have just seen that $h^{p^{n-2}(p-1)} \not\equiv 1 \bmod p^n$. So by Prop. 9.3, $h \in \{g, g+p\}$ is a primitive root mod p^n. □

10. The RSA Cryptosystem

The basic idea of *Public Key Cryptography* is that each participant has two keys: A *public key* that is known to everybody and serves to encrypt messages, and a *private key* that is known only to her or him and is used to decrypt messages. For this idea to work, two conditions have to be satisfied:

(1) Both encryption and decryption must be reasonably fast (with keys of a size satisfying the next condition)
(2) It must be impossible to compute the private key from the public key in less than a very large amount of time (how large will depend on the desired level of security)



The first published system (1977) satisfying these assumptions was designed by Rivest, Shamir and Adleman, and is called the *RSA Cryptosystem* (after their initials). However, already in 1973, Clifford Cocks at GCHQ (the British NSA equivalent) came up with the same system. It was not used by GCHQ, and Cocks' contribution only publicly acknowledged in 1997.

The idea was that finding the prime factors of a large number is very hard, whereas knowing them would allow you to do ceratin things quickly.

10.1. **The set-up.** To generate a public-private key pair, one takes two large prime numbers p and q (of 160 or more decimal digits, say). The public key then consists of $n = pq$ and another positive integer e that has to be coprime with $\operatorname{lcm}(p-1, q-1)$ and can be taken to be fairly (but not too) small (in order to make encryption more efficient).

Encryption proceeds as follows. The message is encoded in one or several numbers $0 \leq m < n$ (e.g., by taking bunches of bits of length less than the length of n (measured in bits)). Then each number m is encrypted as $c = m^e \bmod n$.

In order to decrypt such a c, we need to be able to undo the exponentiation by e. In order to do this, we use Fermat's Little Theorem 8.2 and the Chinese Remainder Theorem 7.1: Since $e \perp \operatorname{lcm}(p-1, q-1)$, we can compute d such that $de \equiv 1 \bmod \operatorname{lcm}(p-1, q-1)$ (using the XGCD Algorithm). Then $c^d = m^{de} \equiv m \bmod p$ and mod$q$ by Fermat's Little Theorem and so $c^d \equiv m \bmod n$ by the Chinese Remainder Theorem.

So the public key is the pair (n, e) and the private key the pair (n, d). Encryption is $m \mapsto m^e \bmod n$, decryption is $c \mapsto c^d \bmod n$.

10.2. **Why is it practical?** Encryption and decryption are reasonably fast: they involve exponentiation mod n, which can be done in $O((\log n)^2 \log e)$ time (where e is the exponent), or even in $O(\log n \log\log n \log e)$, using fast multiplication.

Also, it is possible to select suitable primes p and q in reasonable time: there are algorithms that prove that a given number is prime in polynomial time (polynomial in $\log p$), and gaps between primes are on average of size $\log p$, so one can expect to find a prime in polynomial time. The remaining steps in choosing the public-private key pair are relatively fast.

For example, my laptop running the computer algebra system MAGMA, takes about 4 seconds to find a prime of 100 digits, and about 12 seconds to find a prime of 120 digits.

10.3. **Why is it considered secure?** In order to get m from c, one needs a number t such that $c^t \equiv m \bmod n$. For general m, this means that $te \equiv 1 \bmod \operatorname{lcm}(p-1, q-1)$. Then $te - 1$ is a multiple of $\operatorname{lcm}(p-1, q-1)$, and we can use this in order to factor n in the following way.

Note that if n is an odd prime, then there are exactly two square roots of 1 in the ring $\mathbb{Z}/n\mathbb{Z}$ (which is a field of characteristic not 2 in this case), namely (the residue classes of) 1 and -1. However, when $n = pq$ is the product of two distinct odd primes, then there are four such square roots; they are obtained from pairs of square roots of 1 mod p and mod q via the Chinese Remainder Theorem 7.1. If $x^2 \equiv 1 \bmod n$, but $x \not\equiv \pm 1 \bmod n$, then we can use x to factor n: we have that n divides $x^2 - 1 = (x-1)(x+1)$, but n divides neither factor on the right, so $\gcd(x-1, n)$ has to be a proper divisor of n.

Now suppose we know a multiple f of $\operatorname{lcm}(p-1, q-1)$. Write $f = 2^r s$ with s odd (note that $r \geq 1$ since $p-1$ and $q-1$ are even). Now pick a random $1 < w < n-1$. If $\gcd(w, n) \neq 1$, then we have found a proper divisor of n. Otherwise, we successively compute

$$w_0 = w^s \bmod n\,, w_1 = w_0^2 \bmod n\,, w_2 = w_1^2 \bmod n\,, \ldots\,, w_r = w_{r-1}^2 \bmod n$$

By Fermat's Little Theorem 8.2 and the Chinese Remainder Theorem 7.1, $w_r = 1$. Now if there is some j such that $w_j \not\equiv \pm 1 \bmod n$, but $w_{j+1} \equiv 1 \bmod n$, we have found a square root of 1 mod n that will split n as explained above. One can check (see [Sti, Sect. 5.7.2]) that the probability of success is at least $1/2$. Hence we need no more than two tries on average to factor n.

Conversely, if we have p and q, we can easily compute a suitable t (in fact, our d in the private key is found that way).

The upshot is that in order to break the system, we have to factor n. Now factorization appears to be a hard problem: even though quite some effort has been invested into developing good factoring algorithms (in particular since this is relevant for cryptography! — you can win prize money if you factor certain numbers), and we now have considerably better algorithms than thirty years ago (say), the performance of the best known algorithms is still much worse than polynomial time. The complexity is something like

$$\exp\left(O(\sqrt[3]{\log n (\log\log n)^2})\right).$$

This is already quite a bit better than exponential (in $\log n$), but grows fast enough to make factorization of 300-digit numbers or so infeasible.

For example, again MAGMA on my laptop needs 14 seconds to factor a product of two 20-digit primes and 3 minutes to factor a product of two 30-digit primes.

But note that there is an efficient algorithm (at least in theory) for factoring integers on a quantum computer. So if quantum computers become a reality, cryptosystems based on the difficulty of factorization like RSA will be dead.

11. Discrete Logarithms

In RSA, we use modular exponentiation with a fixed exponent, where the base is the message. There are other cryptosystems, which in some sense work the other way round: they use exponentiation with a fixed base and varying exponent. This can be done in the multiplicative group of a finite field $\mathbb{F}_p$, or even in a more general setting.

11.1. **The Discrete Logarithm Problem.** Let G be a finite cyclic group of order n, with generator g. The problem of finding $a \in \mathbb{Z}/n\mathbb{Z}$ from g and g^a is known as the *Discrete Logarithm Problem*: We want to find the logarithm of g^a to the base g. If $x = g^a$, then sometimes the notation $a = \log_g x$ is used.

The difficulty of this problem depends on the representation of the group G.

(1) The simplest case is $G = \mathbb{Z}/n\mathbb{Z}$ (the additive group), $g = \bar{1}$. Then $\log_g x = x$, and the problem is trivially solved.

(2) It is more interesting to choose $G = \mathbb{F}_p^\times$, with g a primitive root mod p (or rather, its image in $\mathbb{F}_p^\times$). If the group order $\#G = p-1$ has a large prime factor (e.g., $p-1 = 2q$ or $4q$ where q is prime), then here, the DLP

(short for Discrete Logarithm Problem) is considered to be hard. The best known algorithms have complexity

$$O\left(\exp(c\sqrt[3]{\log q(\log\log q)^2})\right) ;$$

the situation is comparable to factorization.

(3) Other groups one can use are the groups of $\mathbb{F}_p$-rational points on *elliptic curves*. Except for certain special cases, no special-purpose algorithms are known, and the best one can do is to use generic algorithms, which have exponential running time $\gg \sqrt{\#G}$. This makes these groups attractive for cryptography, since one gets secure systems with considerably shorter key-lengths.

(4) When the group order $n = \#G$ factors, then the Chinese Remainder Theorem can be used to simplify the problem (if the factorization is known!). (This is the so-called *Pohlig-Hellman attack*.)

11.2. **ElGamal Encryption.** Here is a general setting for a cryptosystem based on DLP. It was originally suggested with $G = \mathbb{F}_p^\times$. In this case, it is advisable to take p such that $p-1$ is a small factor times a (large) prime q, in order to avoid the Pohlig-Hellman attack. Knowing the factorisation of $p-1$ also helps in finding a primitive root g, compare Prop. 9.3 (try random g until one is identified as a primitive root).

It works like this. Bob chooses a random number $a \in \mathbb{Z}/n\mathbb{Z}$, where $n = \#G$ is the group order, and publishes $h = g^a$ as his public key. (The group G and generator g are fixed and also publicly known.) The number a itself is his private key. This means that in order to find the private key from the public key, one has to solve a DLP. Now, when she wants to send Bob a message $m \in G$, Alice also chooses a random number $k \in \mathbb{Z}/n\mathbb{Z}$ and then sends the *pair* (g^k, mh^k) to Bob: she "masks" the message by multiplying it by h^k (remember that h is Bob's public key), but leaves a clue for Bob by also sending g^k. Now to decrypt this, Bob takes the pair (x, y) he receives and computes $m = x^{-a}y$ using his private key a.

An eavesdropper intercepting the ciphertext would need to find $h^k = g^{ak}$ from g^k and $h = g^a$ in order to get the plaintext. This is called the *Diffie-Hellman Problem*, because it also comes up in the secret key exchange protocol developed by Diffie and Hellman (see below). It is believed that the Diffie-Hellman Problem is no easier than the DLP (it is certainly not harder), but this has not been proved.

11.3. **Diffie-Hellman Key Exchange.** This is a method for two people to agree on a secret key, communicating through an open channel. It also works for general cylic groups G with fixed generator g (but was first suggested with $G = \mathbb{F}_p^\times$).

Our two protagonists, Alice and Bob, both select a random number a (for Alice) and b (for Bob) in $\mathbb{Z}/n\mathbb{Z}$. Alice sends $A = g^a$ to Bob, and Bob sends $B = g^b$ to Alice. Then Alice computes $k = B^a$, and Bob computes $k = A^b$. Both get the same result g^{ab}, from which they then can derive a key for a classical symmetric cryptosystem.

In order for the eavesdropper Eve to get at the key, she must be able to find g^{ab} from the knowledge of g^a and g^b, which is exactly the Diffie-Hellman Problem.

12. Quadratic Residues

12.1. **Definition.** Let p be an odd prime and $a \in \mathbb{Z}$ an integer not divisible by p. Then a is called a "quadratic residue mod p" if the congruence $x^2 \equiv a \bmod p$ has solutions. Otherwise, a is a "quadratic nonresidue mod p".

12.2. **Examples.**

p	qu. res.	qu. nonres.
3	1	2
5	1, 4	2, 3
7	1, 2, 4	3, 5, 6
11	1, 3, 4, 5, 9	2, 6, 7, 8, 10

Let g be a primitive root mod p; then each a such that $p \nmid a$ is congruent to some $g^k \bmod p$ (where k is uniquely determined modulo $p-1$, in particular, since p is odd, the parity of k is fixed). It is clear that $x^2 \equiv a \bmod p$ has a solution if and only if k is *even*. Whence:

12.3. **Theorem.** *Let p be an odd prime and $a \in \mathbb{Z}$, $p \nmid a$, and let g be a primitive root mod p. Then the following statements are equivalent.*

(1) *a is a quadratic residue mod p.*
(2) *$\log_g a$ is even.*
(3) *$a^{(p-1)/2} \equiv 1 \bmod p$ (Euler's criterion).*

Proof. We have already seen the equivalence of the first two statements. Now if a is a quadratic residue, then $a \equiv x^2 \bmod p$ for some x, hence $a^{(p-1)/2} \equiv x^{p-1} \equiv 1 \bmod p$ by Fermat's Little Theroem 8.2. On the other hand, if $a^{(p-1)/2} \equiv 1 \bmod p$, then, writing $a \equiv g^k$, the logarithm $k = \log_g a$ cannot be odd, since then $a^{(p-1)/2} \equiv g^{k(p-1)/2} \not\equiv 1 \bmod p$, because $k(p-1)/2$ is not divisible by $p-1$. □

We see that the product of two quadratic residues is again a quadratic residue, whereas the product of a quadratic residue and a quadratic nonresidue is a quadratic nonresidue. Also, the product of two quadratic nonresidues is a quadratic residue.

We also see that there are exactly $(p-1)/2$ quadratic residue classes and $(p-1)/2$ quadratic nonresidue classes mod p.

12.4. **Definition.** To simplify notation, one introduces the *Legendre Symbol*: For p an odd prime and a an integer, set

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{if } p \nmid a \text{ and } a \text{ is a quadratic residue mod } p, \\ -1 & \text{if } p \nmid a \text{ and } a \text{ is a quadratic nonresidue mod } p, \\ 0 & \text{if } p \mid a. \end{cases}$$

By the definitions, we have $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$ if $a \equiv b \bmod p$.

We can combine this definition with Euler's criterion to obtain the following.

12.5. **Proposition.** *Let p be an odd prime, and let $a \in \mathbb{Z}$. Then*

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \bmod p\,,$$

and this congruence determines the value of the Legendre symbol.

Proof. Since $p \geq 3$, the residue classes of -1, 0 and 1 mod p are distinct, so the last statement follows. To prove the congruence, we consider the three possible cases in the definition of the Legendre symbol. If a is a quadratic residue, then both sides are $\equiv 1$ by Thm. 12.3. If a is a quadratic nonresidue, then the left hand side is -1, whereas the right hand side is $\not\equiv 1$, but its square is $\equiv 1$. Since $\mathbb{Z}/p\mathbb{Z}$ is a field, the right hand side must be $\equiv -1$. Finally, if $p \mid a$, then both sides are $\equiv 0$. □

Note that this result tells us that we can determine efficiently whether a given integer a is a quadratic residue mod p or not: the modular exponentiation $a^{(p-1)/2}$ mod p can be computed in polynomial time.

It is a different matter to actually *find* a square root of a mod p if a is a quadratic residue mod p. There are probabilistic polynomial time algorithms for that, but (as far as I know) no *deterministic* polynomial time algorithm is known that works for general p.

12.6. **Theorem.** *For p an odd prime and integers a and b,*

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)\,.$$

Proof. We have

$$\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) \equiv a^{(p-1)/2} b^{(p-1)/2} = (ab)^{(p-1)/2} \equiv \left(\frac{ab}{p}\right) \bmod p$$

by Prop. 12.5. By the same proposition, the value of the Legendre symbol is determined by the congruence. The claim follows. □

12.7. **Example.** By the preceding result, we can compute $\left(\frac{a}{p}\right)$ in terms of the factors of a. So, if $a = \pm 2^e q_1^{f_1} q_2^{f_2} \dots q_k^{f_k}$ with odd primes q_j, then

$$\left(\frac{a}{p}\right) = \left(\frac{\pm 1}{p}\right)\left(\frac{2}{p}\right)^e \left(\frac{q_1}{p}\right)^{f_1} \left(\frac{q_2}{p}\right)^{f_2} \dots \left(\frac{q_k}{p}\right)^{f_k}\,.$$

13. Quadratic Reciprocity

By the preceding example, in order to be able to compute $\left(\frac{a}{p}\right)$ in general, we need to know $\left(\frac{-1}{p}\right)$ and $\left(\frac{2}{p}\right)$, and we need a way to find $\left(\frac{q}{p}\right)$ if $q \neq p$ is another odd prime.

The first is simple.

13.1. **Theorem.** *If p is an odd prime, then*

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2} = \begin{cases} 1 & \text{if } p \equiv 1 \bmod 4, \\ -1 & \text{if } p \equiv 3 \bmod 4. \end{cases}$$

Proof. By Prop. 12.5,

$$\left(\frac{-1}{p}\right) \equiv (-1)^{(p-1)/2} \bmod p\,.$$

Since both sides are ± 1, equality follows. □

So the quadratic character of -1 mod p depends on p mod 4. Is there a similar result concerning the quadratic character of 2 mod p? Here is a table.

p	3	5	7	11	13	17	19	23	29	31
$\left(\frac{2}{p}\right)$	−	−	+	−	−	+	−	+	−	+

It appears that $\left(\frac{2}{p}\right) = 1$ if $p \equiv 1$ or 7 mod 8 and $\left(\frac{2}{p}\right) = -1$ if $p \equiv 3$ or 5 mod 8.

In order to prove a statement like this, we need some other way of expressing the sign of the Legendre symbol. This is provided by the following result due to Gauss.

13.2. **Theorem.** *Let p be an odd prime, and let $S \subset \mathbb{Z}$ be a set of cardinality $(p-1)/2$ such that $\{0\} \cup S \cup -S$ is a complete system of representatives for the residue classes mod p. (Examples are $S = \{1, 2, \ldots, (p-1)/2\}$ and $S = \{2, 4, 6, \ldots, p-1\}$.) Then for all a such that $p \nmid a$, we have*

$$\left(\frac{a}{p}\right) = (-1)^{\#\{s \in S\,:\,\overline{as} \in -\bar{S}\}}\,.$$

Here $\bar{S} = \{\bar{s} : s \in S\}$ is the set of residue classes mod p represented by elements of S.

Proof. For all $s \in S$, there are unique $t(s) \in S$ and $\varepsilon(s) \in \{\pm 1\}$ such that $as \equiv \varepsilon(s)t(s) \bmod p$. We claim that $s \mapsto t(s)$ is a permutation of S. But it is clear that this map is surjective: Let $s \in S$ and b an inverse of a mod p, then there is $s' \in S$ such that $\pm s' \equiv bs \bmod p$, so $as' \equiv \pm s \bmod p$ and therefore $t(s') = s$. So the map must be a bijection.

Now, mod p, we have

$$\begin{aligned}
\left(\frac{a}{p}\right) \prod_{s\in S} s &\equiv a^{(p-1)/2} \prod_{s\in S} s \\
&= \prod_{s\in S} (as) \\
&\equiv \prod_{s\in S} (\varepsilon(s)t(s)) \\
&= \prod_{s\in S} \varepsilon(s) \prod_{s\in S} s \\
&= (-1)^{\#\{s\in S\,:\,\varepsilon(s)=-1\}} \prod_{s\in S} s\,.
\end{aligned}$$

Since p does not divide $\prod_{s\in S} s$, we get

$$\left(\frac{a}{p}\right) \equiv (-1)^{\#\{s\in S\,:\,\varepsilon(s)=-1\}} = (-1)^{\#\{s\in S\,:\,\overline{as}\in -\bar{S}\}} \bmod p\,,$$

and therefore equality (both sides are ± 1). □

Taking $a=-1$ in the preceding result immediately gives Thm. 13.1 again.

We can now use this to prove our conjecture about the value of $\left(\frac{2}{p}\right)$.

13.3. **Theorem.** *If p is an odd prime, then*

$$\left(\frac{2}{p}\right) = (-1)^{(p^2-1)/8} = \begin{cases} 1 & \text{if } p\equiv \pm 1 \bmod 8, \\ -1 & \text{if } p\equiv \pm 3 \bmod 8. \end{cases}$$

Proof. We use Thm. 13.2. For S, we take the standard set

$$S = \{1, 2, 3, \ldots, \frac{p-1}{2}\}\,.$$

We have to count how many elements of S land outside S (mod p) when multiplied by 2.

If $p=8k+1$, these elements are $2k+1, 2k+2, \ldots, 4k$; there are $2k$ of them, an even number, so $\left(\frac{2}{p}\right)=1$.

If $p=8k+3$, these elements are $2k+1, 2k+2, \ldots, 4k+1$; there are $2k+1$ of them, an odd number, so $\left(\frac{2}{p}\right)=-1$.

f $p=8k+5$, these elements are $2k+2, \ldots, 4k+2$; there are $2k+1$ of them, an odd number, so $\left(\frac{2}{p}\right)=-1$.

If $p=8k+7$, these elements are $2k+2, 2k+2, \ldots, 4k+3$; there are $2k+2$ of them, an even number, so $\left(\frac{2}{p}\right)=1$. □

13.4. Do we get similar results for $\left(\frac{q}{p}\right)$, where q is a fixed odd prime and p varies?

Experimental evidence suggests that

$$\left(\frac{3}{p}\right) = \left\{ \begin{array}{rl} 1 & \text{if } p\equiv \pm 1 \bmod 12, \\ -1 & \text{if } p\equiv \pm 5 \bmod 12; \end{array} \right\} = \left\{ \begin{array}{rl} \left(\frac{p}{3}\right) & \text{if } p\equiv 1 \bmod 4, \\ -\left(\frac{p}{3}\right) & \text{if } p\equiv -1 \bmod 4; \end{array} \right.$$

$$\left(\frac{5}{p}\right) = \left\{ \begin{array}{rl} 1 & \text{if } p\equiv \pm 1 \bmod 5, \\ -1 & \text{if } p\equiv \pm 2 \bmod 5. \end{array} \right\} = \left(\frac{p}{5}\right)\,.$$

For larger q, we get similar patterns: if $q\equiv 1 \bmod 4$, the result depends on $p \bmod q$, if $q\equiv -1 \bmod 4$, the result depends on $p \bmod 4q$. Both cases can be combined into the following statement.

13.5. **Theorem (Law of Quadratic Reciprocity).** *Let p and q be distinct odd primes. Then we have*

$$\left(\frac{q}{p}\right) = \left(\frac{p^*}{q}\right) = (-1)^{\frac{p-1}{2}\,\frac{q-1}{2}} \left(\frac{p}{q}\right)$$

$$= \begin{cases} \left(\dfrac{p}{q}\right) & \text{if } p \equiv 1 \bmod 4 \text{ or } q \equiv 1 \bmod 4 \\ -\left(\dfrac{p}{q}\right) & \text{if } p \equiv -1 \bmod 4 \text{ and } q \equiv -1 \bmod 4 \end{cases}$$

where $p^* = (-1)^{(p-1)/2}p$, so $p^* = p$ if $p \equiv 1 \bmod 4$ and $p^* = -p$ if $p \equiv -1 \bmod 4$.

Proof. We make again use of "Gauss' Lemma" Thm. 13.2. We need two sets

$$S = \{1, 2, \ldots, \frac{p-1}{2}\} \quad \text{and} \quad T = \{1, 2, \ldots, \frac{q-1}{2}\}\,.$$

Let $m = \#\{s \in S : \overline{qs} \in -\bar{S}\} \pmod p$ and $n = \#\{t \in T : \overline{pt} \in -\bar{T}\} \pmod q$. Then we have

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^m(-1)^n = (-1)^{m+n}\,.$$

We therefore have to find the parity of the sum $m + n$.

Now, if $qs \equiv -s' \bmod p$ for some $s' \in S$, then there is some $t \in \mathbb{Z}$ such that $pt - qs = s' \in S$, i.e., $0 < pt - qs \le (p-1)/2$. This number t now must be in T:

$$pt > qs > 0 \quad \text{and} \quad pt \le \frac{p-1}{2} + qs \le (q+1)\frac{p-1}{2} < p\frac{q+1}{2}\,.$$

Since q is odd, the last inequality implies $t \le (q-1)/2$. Hence we see that

$$m = \#\{(s,t) \in S \times T : 0 < pt - qs \le \frac{p-1}{2}\}\,.$$

In exactly the same way, we have that

$$n = \#\{(s,t) \in S \times T : -\frac{q-1}{2} \le pt - qs < 0\}\,.$$

Since there is no pair $(s,t) \in S \times T$ such that $pt = qs$, it follows that $m+n = \#X$, where

$$X = \{(s,t) \in S \times T : -\frac{q-1}{2} \le pt - qs \le \frac{p-1}{2}\}\,.$$

This set X is invariant under the rotation by π (or 180°) about the center of the rectangle, which has the effect of changing (s,t) into $(s',t') = (\frac{p+1}{2} - s, \frac{q+1}{2} - t)$:

$$pt' - qs' = p\Big(\frac{q+1}{2} - t\Big) - q\Big(\frac{p+1}{2} - s\Big) = \frac{p-q}{2} - (pt - qs)\,,$$

so $pt - qs \le (p-1)/2 \iff pt' - qs' \ge -(q-1)/2$ and $pt - qs \ge -(q-1)/2 \iff pt' - qs' \le (p-1)/2$. Since the only possible fixed point of the rotation is the center $(\frac{p+1}{4}, \frac{q+1}{4})$ of the rectangle, and since this point belongs to X if it has integral coordinates, we see that

$$\#X \text{ is odd} \iff \frac{p+1}{4}, \frac{q+1}{4} \in \mathbb{Z} \iff p \equiv -1 \bmod 4 \text{ and } q \equiv -1 \bmod 4\,.$$

This concludes the proof. □

13.6. **Example.** With the help of the Law of Quadratic Reciprocity, we can evaluate Legendre symbols in the following way.

$$\left(\frac{67}{109}\right) = \left(\frac{109}{67}\right) = \left(\frac{42}{67}\right) = \left(\frac{2\cdot 3\cdot 7}{67}\right) = \left(\frac{2}{67}\right)\left(\frac{3}{67}\right)\left(\frac{7}{67}\right)$$
$$= (-1)(-\left(\frac{67}{3}\right))(-\left(\frac{67}{7}\right)) = -\left(\frac{1}{3}\right)\left(\frac{4}{7}\right) = -1$$

The disadvantage with this approach is that we have to factor the numbers we get in intermediate stages, which can be very costly if the numbers are large. In order to overcome this difficulty, we generalize the Legendre Symbol and allow arbitrary odd integers instead of odd primes p.

13.7. **Definition.** Let $a \in \mathbb{Z}$, and let n be an odd integer, with factorization $n = \pm p_1^{e_1} p_2^{e_2} \dots p_k^{e_k}$. Then we define the "Jacobi Symbol" via

$$\left(\frac{a}{n}\right) = \prod_{j=1}^{k} \left(\frac{a}{p_j}\right)^{e_j}.$$

It has the following simple properties extending those of the Legendre Symbol.

(1) $\left(\frac{a}{n}\right) = 0$ if and only if $\gcd(a,n) \neq 1$.
(2) If $a \equiv b \bmod n$, then $\left(\frac{a}{n}\right) = \left(\frac{b}{n}\right)$.
(3) $\left(\frac{ab}{n}\right) = \left(\frac{a}{n}\right)\left(\frac{b}{n}\right)$.
(4) $\left(\frac{a}{n}\right) = 1$ if $a \perp n$ and a is a square mod n.

Warning. Contrary to the case of the Legendre symbol (i.e., when n is prime), the converse of the last statement does *not* hold in general. For example, $\left(\frac{2}{15}\right) = 1$, but 2 is not a square mod 15 (since 2 is not a square mod 3 and mod 5).

But, what is more important, the Jacobi symbol also obeys the Law of Quadratic Reciprocity.

13.8. **Theorem.** *Let m and n be positive odd integers. We have*

(1) $\left(\frac{-1}{n}\right) = (-1)^{\frac{n-1}{2}}$.

(2) $\left(\frac{2}{n}\right) = (-1)^{\frac{n^2-1}{8}}$.

(3) $\left(\frac{m}{n}\right) = (-1)^{\frac{m-1}{2}\frac{n-1}{2}}\left(\frac{n}{m}\right)$.

Proof. This is proved by invoking the definition of the Jacobi Symbol and by observing that $n \mapsto (-1)^{(n-1)/2}$ and $n \mapsto (-1)^{(n^2-1)/8}$ are multiplicative on odd integers n, and $(m,n) \mapsto (-1)^{(m-1)(n-1)/4}$ is bimultiplicative on pairs of odd integers. The results then reduce to Thms 13.1, 13.3 and 13.5, respectively. □

13.9. **Example.** Let us compute $\left(\frac{67}{109}\right)$ again.

$$\left(\frac{67}{109}\right) = \left(\frac{109}{67}\right) = \left(\frac{42}{67}\right) = \left(\frac{2}{67}\right)\left(\frac{21}{67}\right) = (-1)\left(\frac{67}{21}\right) = -\left(\frac{4}{21}\right) = -1$$

In general, using Jacobi Symbols in the intermediate steps, we can compute Legendre Symbnols (or, of course, Jacobi Symbols) much in the same way as we compute a GCD; we only have to take care to take out powers of 2 when they appear.

14. Another Proof of Quadratic Reciprocity

Gauss found seven or eight different proofs of the law of quadratic reciprocity in his life. Here is another one, which is more algebraic in flavor, and explains why p^* occurs in a natural way.

14.1. **Definition.** Let p be an odd prime, and set $\zeta = \exp(2\pi i/p) \in \mathbb{C}$. For $a \in \mathbb{Z}$ prime to p, we define the "Gauss Sum"

$$g_a = \sum_{j=1}^{p-1} \left(\frac{j}{p}\right) \zeta^{aj} \in \mathbb{Z}[\zeta]\,.$$

14.2. **Proposition.** *The Gauss Sum has the following properties.*

(1) $g_a = \left(\frac{a}{p}\right) g_1$ *for* $a \perp n$.

(2) $g_1^2 = p^*$.

(3) *For an odd prime* $q \neq p$, *we have* $g_1^q \equiv g_q \bmod q$.
(The congruence takes place in the ring $\mathbb{Z}[\zeta]$.*)*

Proof.

(1) We have

$$\left(\frac{a}{p}\right) g_a = \sum_{j=1}^{p-1} \left(\frac{aj}{p}\right) \zeta^{aj} = \sum_{k=1}^{p-1} \left(\frac{k}{p}\right) \zeta^{k} = g_1\,.$$

(note that $k = aj$ also runs through a complete set of representatives of the primitive residue classes mod p.)

(2) We compute

$$\begin{aligned} g_1^2 &= \sum_{j,k=1}^{p-1} \left(\frac{jk}{p}\right) \zeta^{j+k} = \sum_{j,m=1}^{p-1} \left(\frac{m}{p}\right) \zeta^{j(1+m)} \\ &= \sum_{m=1}^{p-1} \left(\frac{m}{p}\right) \sum_{j=1}^{p-1} \zeta^{j(1+m)} = \left(\frac{-1}{p}\right) p - \sum_{m=1}^{p-1} \left(\frac{m}{p}\right) = p^* \end{aligned}$$

(where $k = jm$; note that $\left(\frac{jk}{p}\right) = \left(\frac{j^2 m}{p}\right) = \left(\frac{m}{p}\right)$. Also note that $\sum_{j=0}^{p-1} \zeta^{ja} = 0$ if $p \nmid a$ and $= p$ otherwise, and that $\sum_{m=1}^{p-1} \left(\frac{m}{p}\right) = 0$.)

(3) Mod q, we have

$$g_1^q = \left(\sum_{j=1}^{p-1} \left(\frac{j}{p}\right) \zeta^{j}\right)^q \equiv \sum_{j=1}^{p-1} \left(\frac{j}{p}\right)^q \zeta^{jq} = \sum_{j=1}^{p-1} \left(\frac{j}{p}\right) \zeta^{qj} = g_q\,.$$

□

14.3. **Remark.** By property (2) above, we have that

$$g_1 = \pm\sqrt{p} \quad \text{if } p \equiv 1 \bmod 4 \qquad \text{and} \qquad g_1 = \pm i\sqrt{p} \quad \text{if } p \equiv 3 \bmod 4.$$

It is then a natural question to ask which of the two signs is the correct one. Gauss was working on this question for quite a long time, until he finally was able to prove that the sign is always "+". (Also for this statement, he found several different proofs in his life.) If $p \equiv 3 \bmod 4$, this has for example the following consequence. It is not hard to see that in this case

$$S(p) = \sum_{a=1}^{p-1} a \left(\frac{a}{p}\right) = -hp$$

with some integer h. The fact that $g_1 = +i\sqrt{p}$ then implies that h is *positive*. This can be interpreted as saying that quadratic nonresidues mod p (between 1 and $p-1$) are larger on average than quadratic residues. (If $p > 3$, then h is the "class number of positive definite binary quadratic forms of discriminant $-p$", which is known to be positive, since it counts something. On the other hand, what one really proves is that

$$h = \frac{ig_1}{p\sqrt{p}} \sum_{a=1}^{p-1} a \left(\frac{a}{p}\right) ,$$

which implies that $S(p) = \mp hp$, and so the sign of the Gauss sum determines the sign of $S(p)$.)

14.4. **Proof of the Quqdratic Reciprocity Law.**

On the one hand, $g_q = \left(\frac{q}{p}\right) g_1$. On the other hand, mod q, we have

$$g_q \equiv g_1^q = g_1 (g_1^2)^{(q-1)/2} = g_1 (p^*)^{(q-1)/2} \equiv g_1 \left(\frac{p^*}{q}\right)$$

(by Euler's criterion). Taking both together, we see that

$$\left(\frac{q}{p}\right) g_1 \equiv \left(\frac{p^*}{q}\right) g_1 \, .$$

Now we multiply by g_1 and use that $g_1^2 = p^*$ is prime to q, so that we can cancel it from both sides. This gives

$$\left(\frac{q}{p}\right) \equiv \left(\frac{p^*}{q}\right) \bmod q$$

and then equality.

15. Sums of Squares

In this section we address the question which positive integers can be written as a sum of two, three, four, ... squares.

Let us first look at sums of two squares. Let

$$\begin{aligned} S &= \{x^2 + y^2 : x, y \in \mathbb{Z}\} \\ &= \{0, 1, 2, 4, 5, 8, 9, 10, 13, 16, 17, 18, 20, 25, 26, 29, 32, 34, 36, 37, 40, \dots\} \, . \end{aligned}$$

It is clear that every square is in S. Also, it is easy to see that if $n \equiv 3 \bmod 4$, then $n \notin S$ (recall that a square is either $\equiv 0$ or $\equiv 1 \bmod 4$). Furthermore, we have the following.

15.1. **Lemma.** *The set S is closed under multiplication: if $m, n \in S$, then $mn \in S$.*

Proof. Note that

$$(x^2 + y^2)(u^2 + v^2) = (xu \mp yv)^2 + (xv \pm yu)^2 .$$

□

What is behind this formula is the following.

$$|x + iy|^2 = x^2 + y^2 \quad \text{and} \quad |\alpha\beta|^2 = |\alpha|^2|\beta|^2 .$$

Because of the multiplicative structure of S, it makes sense to look at the set of prime numbers that are in S. It is clear that $p \notin S$ if $p \equiv 3 \bmod 4$. Obviously, $2 \in S$, and from the list of the first few elements of S, it appears that $p \in S$ if $p \equiv 1 \bmod 4$.

15.2. **Theorem.** *If $p \equiv 1 \bmod 4$ is a prime number, then $p \in S$.*

Proof. We know that -1 is a square mod p, hence there are $a \in \mathbb{Z}$, $k \geq 1$ such that $a^2 + 1 = kp$. We can take $|a| \leq (p-1)/2$, hence we can assume that $k < p/4$.

Now let $k \geq 1$ be minimal such that there are $x, y \in \mathbb{Z}$ with $x^2 + y^2 = kp$. We want to show that $k = 1$. So assume $k > 1$. Let $u \equiv x \bmod k$, $v \equiv y \bmod k$ with $|u|, |v| \leq k/2$. Then

$$u^2 + v^2 = kk'$$

with $1 \leq k' \leq k/2$. (Note that $k' \neq 0$ because $k \nmid p$, as $1 < k < p$.) Now

$$xu + yv \equiv x^2 + y^2 \equiv 0 \bmod k , \qquad xv - yu \equiv xy - yx = 0 \bmod k$$

and $(xu + yv)^2 + (xv - yu)^2 = (x^2 + y^2)(u^2 + v^2) = k^2\, k'p$. If we let

$$x' = \frac{xu + yv}{k} , \qquad y' = \frac{xv - yu}{k} ,$$

then $(x')^2 + (y')^2 = k'p$ and $k' < k$, contradicting our choice of k. So we must have had $k = 1$. □

The technique of proof use here is called "descent" and goes back to Fermat. The name comes from the fact that we "descend" from one value of k to a smaller one.

By what we know so far, we have already proved one direction of the following result characterizing the elements of S.

15.3. **Theorem.** *A positive integer n can be represented as a sum of two squares if and only if for every prime $p \mid n$ with $p \equiv 3 \bmod 4$, the exponent with which p appears in the factorization of n is even.*

Proof. If n is of the specified form, then $n = p_1 \cdots p_r m^2$ with primes $p_j = 2$ or $p_j \equiv 1 \bmod 4$. Since by the above, all factors in this product are in S and S is closed under multiplication, $n \in S$.

Now assume that $n \in S$ and that we already know that all $m \in S$ with $m < n$ are of the specified form. Let $p \equiv 3 \bmod 4$ be a prime number dividing n. Write $n = x^2 + y^2$. We claim that p divides both x and y. It then follows that $n = p^2 m$ with $m = (x/p)^2 + (y/p)^2 \in S$, so we are done by induction.

To show that p divides x and y, assume that (for example), p does not divide x. Then there is $a \in \mathbb{Z}$ with $ax \equiv 1 \bmod p$, and we get

$$0 \equiv a^2 n = (ax)^2 + (ay)^2 \equiv 1 + (ay)^2 \bmod p\,,$$

contradicting the fact that -1 is not a square mod p. So p must divide x and y. $\square$

For three squares, the criterion is simpler (but we will not prove it).

15.4. **Theorem.** *A positive integer can be represented as a sum of three squares if and only if it is not of the form $4^k m$ where $m \equiv 7 \bmod 8$.*

It is easy to see that a number $n = 4^k m$ with $m \equiv 7 \bmod 8$ is not a sum of three squares. First note that if a sum $x^2 + y^2 + z^2$ is divisible by 4, then x, y, z have to be even. This implies that $4n$ is a sum of three squares if and only if n is. So we can assume that $k = 0$. Finally, mod 8, a square is 0, 1 or 4, so a sum of three squares can never be $\equiv 7 \bmod 8$. The hard part of the proof is to show that every n not of the given form actually is a sum of three squares. Part of the difficulty comes from the fact that the set of sums of three squares is *not* closed under multiplication: 3 and 5 are sums of three squares, but 15 is not.

15.5. **Four Squares.** It might therefore seem rather hopeless to look for an identity for four squares analogous to

$$(xu \mp yv)^2 + (xv \pm yu)^2 = (x^2 + y^2)(u^2 + v^2)\,,$$

but in fact there is a good reason why one exists. The *quaternion algebra*, a 4-dimensional $\mathbb{R}$-algebra, is a beautiful analog of the 2-dimensional algebra $\mathbb{C}$; it was discovered by Hamilton. It is defined to be

$$\mathbb{H} := \{a + ib + cj + dk : a, b, c, d \in \mathbb{R}\}\,,$$

with the noncommutative multiplication rules

$$i^2 = j^2 = k^2 = -1,$$
$$ij = k,\ ji = -k, \quad jk = i,\ kj = -i, \quad ki = j,\ ik = -j\,.$$

One can then define a "norm" map

$$\mathrm{N}(a + ib + cj + dk) := (a + ib + cj + dk)(a - ib - cj - dk) = a^2 + b^2 + c^2 + d^2\,,$$

and it is easy to check that the norm is multiplicative. When one writes out what this means, one discovers the identity

$$\begin{aligned}(a^2 + b^2 + c^2 + d^2)&(A^2 + B^2 + C^2 + D^2)\\ &= (aA - bB - cC - dD)^2 + (aB + bA + cD - dC)^2\\ &\quad + (aC + cA - bD + dB)^2 + (aD + dA + bC - cB)^2\,.\end{aligned}$$

In light of this, the set of integers representable by four squares must be closed under multiplication. In fact:

15.6. **Theorem (Lagrange).** *All positive integers are sums of four squares.*

Proof. By the identity stated above, it suffices to show that all primes p are sums of four squares. We do this by descent, imitating the proof of the Two Squares Theorem. First note that, applying Lemma 15.7 below, we can find integers a, b, c, d and k such that $a^2+b^2+c^2+d^2 = kp$ and $1 \le k < p$ (taking $d = 0$, say). If $k = 1$ we are done. Otherwise, let A, B, C, D be the integers determined by

$$\begin{aligned} A &\equiv -a \bmod k, & |A| &\le k/2 \\ B &\equiv b \bmod k, & |B| &\le k/2 \\ C &\equiv c \bmod k, & |C| &\le k/2 \\ D &\equiv d \bmod k, & |D| &\le k/2 \end{aligned}$$

Thus $A^2+B^2+C^2+D^2 \le k^2$. If equality holds, A, B, C and D must each equal $k/2$ or $-k/2$. In that case a, b, c and d are each congruent to $k/2$ modulo k, which means k^2 divides $a^2+b^2+c^2+d^2 = kp$. But that is not possible because $1 < k < p$ and p is prime. Hence $A^2+B^2+C^2+D^2 = kk'$ with $1 \le k' < k$. Applying the magic identity, we have

$$\begin{aligned} k^2k'p &= (a^2+b^2+c^2+d^2)(A^2+B^2+C^2+D^2) \\ &= (aA-bB-cC-dD)^2 + (aB+bA+cD-dC)^2 \\ &\quad + (aC+cA-bD+dB)^2 + (aD+dA+bC-cB)^2 . \end{aligned}$$

Consider the right hand side: the latter three terms, and hence all four terms, are divisible by k^2. Dividing both sides by k^2, we obtain a representation of $k'p$ as a sum of four squares, which completes one step of the descent. As already noted, at each step we have $1 \le k' < k$. So, after a finite number of steps of the descent we must obtain $k' = 1$. This completes the proof. □

15.7. **Lemma.** *Let p be an odd prime. Then there are integers u, v such that $u^2+v^2+1 \equiv 0 \bmod p$.*

Proof. The statement is equivalent to the following: there are $\bar{u}, \bar{v} \in \mathbb{F}_p$ such that $\bar{u}^2 = -\bar{v}^2 - 1$. Now let

$$X = \{\bar{u}^2 : \bar{u} \in \mathbb{F}_p\} \quad \text{and} \quad Y = \{-\bar{v}^2 - 1 : \bar{v} \in \mathbb{F}_p\},$$

then $\#X = \#Y = (p+1)/2$. Since $\#X + \#Y = p+1 > p = \#\mathbb{F}_p$, X and Y cannot be disjoint, which proves the claim. □

16. Geometry of Numbers

In this section, we will learn about a nice method to solve number theoretical problems using geometry. The main result was discovered by Hermann Minkowski. The basic idea is that if we have a sufficiently nice and sufficiently "large" set in $\mathbb{R}^n$, then it will contain a non-zero point with integral coordinates. For the applications, it is convenient to use more general "lattices" than the integral points, so we have to introduce this notion first.

16.1. **Definition.** A *lattice* $\Lambda \subset \mathbb{R}^n$ is the set of all integral linear combinations of a set of basis vectors $v_1, \dots, v_n$ of $\mathbb{R}^n$. In particular, Λ is a subgroup of the additive group $\mathbb{R}^n$. The set

$$F = \Big\{ \sum_{j=1}^{n} t_j v_j : 0 \le t_j < 1 \text{ for all } j \Big\}$$

is called a *fundamental parallelotope* for Λ. $\Delta(\Lambda) = \mathrm{vol}(F) = |\det(v_1, \dots, v_n)|$ is the *covolume* of Λ.

The most important property of F is that every vector $v \in \mathbb{R}^n$ can be written *uniquely* as $v = \lambda + w$ with $\lambda \in \Lambda$ and $w \in F$. In other words, $\mathbb{R}^n$ is the disjoint union of all translates $F + \lambda$ of F by elements of Λ.

16.2. **Example.** The standard example of a lattice is $\Lambda = \mathbb{Z}^n \subset \mathbb{R}^n$, which is generated by the standard basis $e_1, \dots, e_n$ of $\mathbb{R}^n$ and has covolume $\Delta(\mathbb{Z}^n) = 1$.

In some sense, this is the only example: if $\Lambda = \mathbb{Z}v_1 + \dots + \mathbb{Z}v_n \subset \mathbb{R}^n$ is any lattice, then Λ is the image of $\mathbb{Z}^n$ under the invertible linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ that sends e_j to v_j. The covolume $\Delta(\Lambda)$ is then $|\det(T)|$.

16.3. **Proposition.** *Let $\Lambda \subset \mathbb{R}^n$ be a lattice, and let $\Lambda' \subset \Lambda$ be a subgroup of finite index m. Then Λ' is also a lattice, and $\Delta(\Lambda') = m\,\Delta(\Lambda)$.*

Proof. As an abstract abelian group, $\Lambda \cong \mathbb{Z}^n$. By the structure theorem for finitely generated abelian groups, there is an isomorphism $\phi : \Lambda \to \mathbb{Z}^n$ that sends Λ' to $a_1\mathbb{Z} \times \dots \times a_n\mathbb{Z}$ with nonnegative integers $a_1, \dots, a_n$. Since the index m of Λ' in Λ is finite, we have $a_1 \cdots a_n = m$. Let $v_1, \dots, v_n$ be the generators of Λ that are sent to the standard basis of $\mathbb{Z}^n$ under ϕ. Then $\Lambda' = \mathbb{Z}a_1v_1 + \dots + \mathbb{Z}a_nv_n \subset \mathbb{R}^n$, so Λ' is a lattice. Furthermore,

$$\Delta(\Lambda') = |\det(a_1v_1, \dots, a_nv_n)| = a_1 \cdots a_n\, |\det(v_1, \dots, v_n)| = m\,\Delta(\Lambda)\,.$$

□

16.4. **Corollary.** *Let $\phi : \mathbb{Z}^n \to M$ be a group homomorphism* onto *a finite group M. Then the kernel of ϕ is a lattice Λ, and $\Delta(\Lambda) = \#M$.*

Proof. By the standard isomorphism theorem, we have $\mathbb{Z}^n/\ker\phi \cong M$, hence $\Lambda = \ker\phi$ is a subgroup of the lattice $\mathbb{Z}^n$ of finite index $\#M$. The claim follows from Prop. 16.3 and $\Delta(\mathbb{Z}^n) = 1$. □

Now we are ready to state and prove Minkowski's Theorem.

16.5. **Theorem (Minkowski).** *Let $\Lambda \subset \mathbb{R}^n$ be a lattice, and let $S \subset \mathbb{R}^n$ be a* symmetric *(i.e., $S = -S$) and* convex *subset such that* $\mathrm{vol}(S) > 2^n\Delta(\Lambda)$. *Then S contains a nonzero lattice point from Λ.*

Proof. In a first step, we show that $X = \frac{1}{2}S$ has to intersect one of its translates under elements of Λ. Let F be a fundamental parallelotope for Λ, and for $\lambda \in \Lambda$, set

$$X_\lambda = F \cap (X + \lambda)\,.$$

By the fundamental property of F, we get that

$$X = \coprod_{\lambda \in \Lambda} (X_\lambda - \lambda)$$

(i.e., X is a *disjoint* union of translates of the X_λ). Hence

$$\sum_\lambda \operatorname{vol}(X_\lambda) = \operatorname{vol}(X) = 2^{-n} \operatorname{vol}(S) > \Delta(\Lambda) = \operatorname{vol}(F)\,,$$

and so the sets X_λ cannot be all disjoint (because they then would not fit into F). So there are $\lambda \neq \mu$ such that $X_\lambda \cap X_\mu \neq \emptyset$. Shifting by $-\mu$, we see that

$$X \cap (X + \lambda - \mu) \neq \emptyset\,.$$

Let x be a point in the intersection. Then $2x \in S$ and $2x - 2(\lambda - \mu) \in S$. Since S is symmetric, we also have $2(\lambda - \mu) - 2x \in S$. Then, since S is also convex, the midpoint of the linesegment joining $2x$ and $2(\lambda - \mu) - 2x \in S$ must also be in S. But this midpoint is $\lambda - \mu \in \Lambda \setminus \{0\}$, and the statement is proved. □

Let us use this result to re-prove the essential results on sums of two and four squares.

16.6. **Theorem.** *Let $p \equiv 1 \bmod 4$ be a prime. Then p is a sum of two squares.*

Proof. We need a lattice Λ and a set S. Let u be a square root of $-1 \bmod p$, and set

$$\Lambda = \{(x, y) \in \mathbb{Z}^2 : x \equiv uy \bmod p\}\,.$$

Then Λ is a lattice in $\mathbb{R}^2$, and $\Delta(\Lambda) = p$ (we can think of Λ as the kernel of the composition

$$\mathbb{Z}^2 \longrightarrow \mathbb{F}_p^2 \longrightarrow \frac{\mathbb{F}_p^2}{\langle(\bar{u}, 1)\rangle}$$

which is a surjective group homomorphism onto a group of order p). For the set S, we take the open disk

$$S = \{(\xi, \eta) \in \mathbb{R}^2 : \xi^2 + \eta^2 < 2p\}\,.$$

Then $\operatorname{vol}(S) = \pi \cdot 2p = 2\pi p > 4p = 2^2\Delta(\Lambda)$, and so by Thm. 16.5, there is some nonzero $(x, y) \in \Lambda \cap S$. Now for each $(x, y) \in \Lambda$, we have that

$$x^2 + y^2 \equiv (uy)^2 + y^2 = y^2(1 + u^2) \equiv 0 \bmod p\,.$$

So p divides $x^2 + y^2$; on the other hand, $0 < x^2 + y^2 < 2p$ by the definition of S. So we must have $x^2 + y^2 = p$. □

Now let us consider the case of four squares. From Lemma 15.7, we know that for every odd prime p, there are integers u and v such that p divides $1 + u^2 + v^2$.

16.7. **Theorem.** *Let p be an odd prime. Then p is a sum of four squares.*

Proof. We need again a lattice Λ and a set S. For S, we should obviously take a suitable open ball:

$$S = \{(\xi_1, \xi_2, \xi_3, \xi_4) \in \mathbb{R}^4 : \xi_1^2 + \xi_2^2 + \xi_3^2 + \xi_4^2 < 2p\}\,.$$

What is the volume of S? Here it is useful to know the general formula for the volume of the n-dimensional unit ball; it is

$$\operatorname{vol}(B^n) = \frac{\pi^{n/2}}{\left(\frac{n}{2}\right)!}$$

(where for odd n, the factorial satisfies the usual recurrence $(x+1)! = x!\,(x+1)$, and one has $(-1/2)! = \sqrt{\pi}$). For $n = 4$, we get $\pi^2/2$ for the volume of the unit ball, hence $\operatorname{vol}(S) = \pi^2(2p)^2/2 = 2\pi^2 p^2$.

From this, we can already see that the lattice should have covolume p^2. This means that we need a 2-dimensional subspace of $\mathbb{F}_p^4$ on which $x_1^2 + x_2^2 + x_3^2 + x_4^2$ vanishes. One such subspace is given by

$$V = \langle (1, \bar{u}, \bar{v}, 0), (0, -\bar{v}, \bar{u}, 1) \rangle \ :$$

if $(\bar{a}, \bar{a}\bar{u} - \bar{b}\bar{v}, \bar{a}\bar{v} + \bar{b}\bar{u}, \bar{b})$ is a general element of V, then

$$\begin{aligned} \bar{a}^2 + (\bar{a}\bar{u} - \bar{b}\bar{v})^2 + (\bar{a}\bar{v} + \bar{b}\bar{u})^2 + \bar{b}^2 \\ &= \bar{a}^2(1 + \bar{u}^2 + \bar{v}^2) + \bar{b}^2(\bar{v}^2 + \bar{u}^2 + 1) + 2\bar{a}\bar{b}(-\bar{u}\bar{v} + \bar{v}\bar{u}) \\ &= 0\,. \end{aligned}$$

If Λ is the kernel of $\mathbb{Z}^4 \to \mathbb{F}_p^4 \to \mathbb{F}_p^4/V$, then for $(x_1, x_2, x_3, x_4) \in \Lambda$, we have $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4) \in V$, hence p divides $x_1^2 + x_2^2 + x_3^2 + x_4^2$. For the covolume, we have $\Delta(\Lambda) = \#(\mathbb{F}_p^4/V) = p^2$. Since $\mathrm{vol}(S) = 2\pi^2 p > 16p^2$, the proof can be concluded in the same way as before. □

17. Ternary Quadratic Forms

In the preceding sections, we have seen some *quadratic forms*.

17.1. **Definition.** An *n-ary quadratic form* is a homogenous polynomial of degree 2 in n variables (here, the coefficients will always be integers, but one can consider quadratic forms over any ring). For $n = 2$, we have *binary* quadratic forms; they have the general form

$$Q(x, y) = a\,x^2 + b\,xy + c\,y^2\,.$$

For $n = 3$, we have *ternary* quadratic forms

$$Q(x, y, z) = a\,x^2 + b\,y^2 + c\,z^2 + d\,xy + e\,yz + f\,zx\,,$$

and so on.

So far, we have been asking about *representations* of numbers by a quadratic form Q, i.e., whether it is possible to find a given integer as the value of Q at some tuple of integers.

Another question one can ask is whether a given quadratic form has a nontrivial zero, i.e., whether there exist (in the case of ternary forms, say) integers x, y, z, not all zero, such that $Q(x, y, z) = 0$. This is what we will look into now. For binary forms, this question is not very interesting; it boils down to deciding whether or not the form is the product of two linear forms with integral coefficients, which is the case if and only if the *discriminant* $b^2 - 4ac$ of the form is a square. For ternary forms, however, this is an interesting problem. Note that we can always assume that solutions are *primitive* (i.e., $\gcd(x, y, z) = 1$): common divisors can always be divided out.

17.2. **Definition.** Let Q be a quadratic form in n variables; then it can be given by a symmetric matrix M_Q whose off-diagonal entries can be half-integers, such that $Q(\mathbf{x}) = \mathbf{x}^\top M_Q \mathbf{x}$. Then $\det Q = \det(M_Q)$ is called the *determinant* of Q, and $\operatorname{disc} Q = (-1)^{n-1} 2^{2\lfloor n/2 \rfloor} \det Q$ is called the *discriminant* of Q; the discriminant is always an integer. (The reason for the power of 2 appearing in the definition of $\operatorname{disc}(Q)$ is that the discriminant then also makes sense in characteristic 2.)

For example,

$$\operatorname{disc}(a\,x^2 + b\,xy + c\,y^2) = b^2 - 4\,ac$$

and

$$\operatorname{disc}(a\,x^2 + b\,y^2 + c\,z^2 + d\,xy + e\,yz + f\,zx) = 4\,abc + def - ae^2 - bf^2 - cd^2\,.$$

A quadratic form Q is *non-degenerate* if $\operatorname{disc} Q \neq 0$, otherwise it is called *degenerate* or *singular*. In this latter case, there is a linear transformation of the variables that results in a quadratic form involving fewer variables (choose an element in the kernel of M_Q as one of the new basis vectors ...).

17.3. **Some Geometry.** Ternary quadratic forms correspond to *conic sections* in the plane. If we are looking for solutions to $Q(x, y, z) = 0$ in real numbers such that $z \neq 0$ (say), we can divide by z^2 and set $\xi = x/z$, $\eta = y/z$ to obtain $Q(\xi, \eta, 1) = 0$, the equation of a conic section in $\mathbb{R}^2$. (If we want to include the solutions with $z = 0$, we have to consider the conic section in the *projective* plane.) In this setting, nontrivial primitive integral solutions to $Q(x, y, z) = 0$ correspond to *rational points* (points with rational coordinates) on the conic. This correspondence is two-to-one: to the point $(x/z, y/z)$ (in lowest terms) there correspond the two solutions (x, y, z) and $(-x, -y, -z)$.

For example, if we take $Q(x, y, z) = x^2 + y^2 - z^2$, then it corresponds to the unit circle in the xy plane, and the solutions (in this case, Pythagorean Triples) correspond to the rational points on the unit circle (there are no solutions with $z = 0$). In fact, it is easy to describe them all: fix one point, say $(-1, 0)$, and draw a line with rational slope $t = u/v$ through it. It will intersect the circle in another point, which will again have rational coordinates. Conversely, if we take some rational point on the circle, the line connecting it to $(-1, 0)$ will have rational slope. We see that the rational points are parametrized by the rational slopes (including $\infty = 1/0$ for the vertical tangent at $(-1, 0)$; this line gives $(-1, 0)$ itself). The same kind of argument can be used quite generally.

17.4. **Theorem.** *Let $Q(x, y, z)$ be a non-degenerate ternary quadratic form that has a primitive integral solution (x_0, y_0, z_0). Then there are binary quadratic forms R_x, R_y and R_z such that, up to scaling, all integral solutions of $Q(x, y, z) = 0$ are given by*

$$(R_x(u, v), R_y(u, v), R_z(u, v))$$

with integers u, v.

Proof. We first assume that $Q = y^2 - xz$. Then we can clearly take

$$R_x(u, v) = u^2\,, \quad R_y(u, v) = uv\,, \quad R_z(u, v) = v^2\,.$$

(Dividing by z^2, we have $(y/z)^2 = x/z$; put $y/z = u/v$ and clear denominators.)

Now assume that $(x_0, y_0, z_0) = (1, 0, 0)$. Then

$$Q(x, y, z) = b\,y^2 + c\,z^2 + d\,xy + e\,yz + f\,zx\,.$$

If we set

$$x = b\,X + e\,Y + c\,Z\,, \quad y = -d\,X - f\,Y\,, \quad z = -d\,Y - f\,Z\,,$$

then $Q(x, y, z) = -\operatorname{disc}(Q)(Y^2 - XZ)$, as is easily checked. By the first case (note that $\operatorname{disc}(Q) \neq 0$), this means that

$$R_x(u, v) = b\,u^2 + e\,uv + c\,v^2\,, \quad R_y(u, v) = -d\,u^2 - f\,uv\,, \quad R_z(u, v) = -d\,uv - f\,v^2$$

do what we want.

Finally, we consider the general case. By Prop. 18.6 in the "Introductory Algebra" notes (Fall 2005), there is a matrix $T \in \mathrm{GL}_3(\mathbb{Z})$ such that $(x_0\ y_0\ z_0) = (1\ 0\ 0)T$. Write

$$(x\ y\ z) = (x'\ y'\ z')T$$

and set $Q'(x', y', z') = Q(x, y, z)$; then $Q'(1, 0, 0) = Q(x_0, y_0, z_0) = 0$. By the previous case, we have binary quadratic forms R'_x, R'_y, R'_z that parametrize the solutions of Q'. Then

$$(R_x\ R_y\ R_z) = (R'_x\ R'_y\ R'_z)T$$

are the binary quadratic forms we want for Q. □

17.5. **Example.** For $Q(x, y, z) = x^2 + y^2 - z^2$ and the initial solution $(-1, 0, 1)$, we can choose

$$T = \begin{pmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and obtain $x = -x'$, $y = y'$, $z = x' + z'$, so

$$Q'(x', y', z') = Q(-x', y', x' + z') = (y')^2 - (z')^2 - 2x'z' \,.$$

The quadratic forms parametrizing the solutions of Q' are

$$R'_x(u, v) = u^2 - v^2 \,, \quad R'_y(u, v) = 2\,uv \,, \quad R'_z(u, v) = 2\,v^2 \,.$$

For our original form Q, we then get

$$\begin{aligned} R_x(u, v) &= -R'_x(u, v) = v^2 - u^2 \\ R_y(u, v) &= R'_y(u, v) = 2\,uv \\ R_z(u, v) &= R'_x(u, v) + R'_z(u, v) = u^2 + v^2 \end{aligned}$$

This is exactly the well-known parametrization of the Pythagorean Triples.

We see that we can easily find all solutions if we know just one. So there are two questions that remain: to decide whether a solution exist, and, if so, find one.

18. Legendre's Theorem

We can always *diagonalize* a non-degenerate quadratic form by a suitable linear substitution of the variables (and perhaps scaling, to keep the coefficients integral). Basically, this comes down to repeatedly completing the square. So, for theoretical purposes at least, we can assume that our ternary quadratic form is *diagonal*:

$$Q(x, y, z) = a\,x^2 + b\,y^2 + c\,z^2 \,.$$

In practice, it might be a very bad idea to do this, as the coefficients a, b, c may be much larger than the coefficients of the original form!

Let us be a bit more formal.

18.1. **Definition.** Let Q, Q' be two ternary quadratic forms. We say that Q and Q' are *equivalent* if

$$Q'(x,y,z) = \lambda\, Q(a_{11}x + a_{12}y + a_{13}z, a_{21}x + a_{22}y + a_{23}z, a_{31}x + a_{32}y + a_{33}z)$$

with $\lambda \in \mathbb{Q}^\times$ and a matrix

$$T = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \in \mathrm{GL}_3(\mathbb{Q})\,.$$

The above then means that every non-degenerate ternary quadratic form is equivalent to a diagonal one. It is easily seen that Q has nontrivial integral (or equivalently, rational) solutions if and only if Q' does.

If we want to decide whether $Q = ax^2+by^2+cz^2$ admits a solution, we can simplify the problem somewhat. We can, of course, assume that $\gcd(a,b,c) = 1$. If a (say) is divisible by a square d^2, then we can as well move d^2 into the x^2 term and thus obtain an equivalent form with smaller coefficients. Proceeding in this way, we can assume that a, b and c are squarefree.

Also, if two of the coefficients, say b and c, have a common prime divisor p, then p must divide x. We replace x by px and then divide the form by p, making the coefficients smaller. In this way, we can also assume that a, b and c are coprime in pairs. We can summarize these assumptions by saying that abc is squarefree.

18.2. **Necessary Conditions for Solubility.** We can easily write down a number of conditions that are necessary for the existence of a solution:

(1) Not all of a, b, and c have the same sign.
(2) If abc is odd, then a, b and c are not equal mod 4.
(3) If a is even (say), then $b + c \equiv 0$ or $a + b + c \equiv 0 \bmod 8$.
(4) If $p \mid a$ is odd, then $-bc$ is a quadratic residue mod p.
(5) If $p \mid b$ is odd, then $-ca$ is a quadratic residue mod p.
(6) If $p \mid c$ is odd, then $-ab$ is a quadratic residue mod p.

For odd primes p such that $p \nmid abc$, we do not obtain any restrictions in this way: there are always nontrivial solutions mod p (compare Lemma 15.7; the proof is more or less the same).

Note that in order to check the conditions, we have to factor the coefficients a, b and c. It can be shown that this cannot be avoided: if one can find solutions to (diagonal) ternary quadratic forms, then one can also factor integers, hence solving ternary quadratic forms is at least as hard as factoring integers.

The surprising fact is that these necessary conditions are already sufficient!

18.3. **Theorem (Legendre).** *Let $Q(x,y,z) = a\,x^2+b\,y^2+c\,z^2$ with abc squarefree satisfy the conditions in 18.2. Then there exists a nontrivial solution in integers.*

Proof. We will prove this using Minkowski's Theorem 16.5. Let $D = |abc|$. Our first claim is that there is a lattice $\Lambda \subset \mathbb{Z}^3$ such that for all $(x,y,z) \in \Lambda$, $2D$ divides $Q(x,y,z)$, and such that $\Delta(\Lambda) = 2D$. In order to find such a Λ, we construct lattices Λ_p for all odd $p \mid D$ such that $p \mid Q(x,y,z)$ when $(x,y,z) \in \Lambda_p$ and such that $\Delta(\Lambda_p) = p$. We will also construct a lattice Λ_2 such that 2 or 4 divides $Q(x,y,z)$ for all $(x,y,z) \in \Lambda_2$ (according to whether abc is odd or even) and such that $\Delta(\Lambda_2) = 2$ or 4. Then $\Lambda = \bigcap_{p\mid D} \Lambda_p$ will do what we want.

Now let p be an odd prime divisor of a (similarly for b or c). By assumption, there exists some $u_p \in \mathbb{Z}$ such that p divides $bu_p^2 + c$. Let

$$\Lambda_p = \{(x, y, z) \in \mathbb{Z}^3 : y \equiv u_p z \bmod p\}\,.$$

It is easily checked that Λ_p does what we want.

Now assume that abc is odd. Then we let

$$\Lambda_2 = \{(x, y, z) \in \mathbb{Z}^3 : x + y + z \equiv 0 \bmod 2\}\,.$$

If a (say) is even and $b + c \equiv 0 \bmod 4$, we let

$$\Lambda_2 = \{(x, y, z) \in \mathbb{Z}^3 : x \equiv y + z \equiv 0 \bmod 2\}\,;$$

if $b + c \equiv 2 \bmod 4$, we let

$$\Lambda_2 = \{(x, y, z) \in \mathbb{Z}^3 : x \equiv y \equiv z \bmod 2\}\,.$$

It is again easily checked that Λ_2 has the required properties in each case.

Now assume that the sign of c differs from that of a and b. Then we take for S the elliptic cylinder

$$S = \{(\xi, \eta, \zeta) \in \mathbb{R}^3 : |a|\xi^2 + |b|\eta^2 < 2D \text{ and } |c|\zeta^2 < 2D\}\,.$$

We find that

$$\mathrm{vol}(S) = \pi \frac{2D}{\sqrt{|ab|}} 2 \frac{\sqrt{2D}}{\sqrt{|c|}} = \frac{4\sqrt{2}\pi D\sqrt{D}}{\sqrt{D}} = 4\sqrt{2}\pi D > 16D = 8\Delta(\Lambda)\,.$$

Hence by Thm. 16.5, there is a nonzero element (x, y, z) in $\Lambda \cap S$. Since it is in Λ, $Q(x, y, z)$ is a multiple of $2D$. Since $|Q(x, y, z)| = \big|(|a|x^2 + |b|y^2) - |c|z^2\big|$ and both terms in the difference are $< 2D$, we find that $|Q(x, y, z)| < 2D$. Together, these imply that $Q(x, y, z) = 0$. □

Note that the ellipsoid given by $|a|\xi^2 + |b|\eta^2 + |c|\zeta^2 < 2D$ would be too small for the proof to work. Note also that we did not need to assume that solutions mod 4 or mod 8 exist. This is a general feature: one can always leave out one "place" in the conditions — either conditions mod powers of 2, or some odd prime, or the "infinite place", which here gives rise to the condition on the signs. The resaon behind this is essentially quadratic reciprocity, which leads to the fact that the number of places where the conditions fail is always *even*. In the above proof, one could use the mod 4/mod 8 conditions to come up with a lattice of covolume $4D$ giving divisibility by $4D$; then the ellipsoid would be sufficiently large, and we need not require the sign condition on the coefficients!

There is also a proof by descent (in fact, that was how Legendre originally proved his theorem).

18.4. **Corollary.** *If $a\,x^2 + b\,y^2 + c\,z^2 = 0$ has a nontrivial solution in integers, then it has one such that*

$$\max\{|a|\,x^2\,, |b|\,y^2\,, |c|\,z^2\} \le 4\pi^{-2/3}|abc| < 1.865|abc|\,,$$

or equivalently,

$$|x| \le 2\pi^{-1/3}\sqrt{|bc|}\,, \quad |y| \le 2\pi^{-1/3}\sqrt{|ca|}\,, \quad |z| \le 2\pi^{-1/3}\sqrt{|ab|}\,.$$

Proof. With $2|abc|$ instead of $4\pi^{-2/3}|abc|$, this follows from the preceding proof. Now note that this proof will still work if in the definition of S, we replace $2D$ by αD with $\alpha > 4\pi^{-2/3}$. Since S contains only finitely many lattice points, there is one solution such that

$$\max\{|a|\,x^2\,,|b|\,y^2\,,|c|\,z^2\} < \alpha|abc|$$

for all $\alpha > 4\pi^{-2/3}$, which implies the claim. □

In fact, more is true.

18.5. **Theorem (Holzer).** *If $a\,x^2 + b\,y^2 + c\,z^2 = 0$ (with abc squarefree) has a nontrivial solution in integers, then it has one such that*

$$\max\{|a|\,x^2\,,|b|\,y^2\,,|c|\,z^2\} \le |abc|\,,$$

or equivalently,

$$|x| \le \sqrt{|bc|}\,, \quad |y| \le \sqrt{|ca|}\,, \quad |z| \le \sqrt{|ab|}\,.$$

To get this (when $a, b > 0$ and $c < 0$, say), one assumes that a given solution has $|z| > \sqrt{ab}$ and constructs a new one from this with smaller $|z|$. So the solution with smallest $|z|$ must have $|z| \le \sqrt{ab}$; the bounds on x and y then follow.

19. p-adic Numbers

19.1. **Motivation.** In many circumstances, one wants to consider statements for all powers of some prime number p. For example, if a polynomial equation has (nontrivial) integral solutions, it necessarily has (nontrivial) solutions modulo all powers of p. We also considered (nontrivial) solutions in real numbers. Now $\mathbb{R}$ is a field, but $\mathbb{Z}/p^n\mathbb{Z}$ is only a ring (finite, which is nice) and not even an integral domain when $n \ge 2$. Therefore it is desirable to work instead in a structure that is an integral domain or a field and at the same time captures statements about all powers of p simultaneously. This can be done by "passing to the limit" in a suitable way and leads to the ring $\mathbb{Z}_p$ of p-adic integers and the field $\mathbb{Q}_p$ of p-adic numbers. Our statement about nontrivial solutions mod p^n for all n can then simply be expressed by saying that there is a (nontrivial) solution in $\mathbb{Z}_p$ (or in $\mathbb{Q}_p$).

Consider, for example, the equation $x^2 + 7 = 0$ modulo powers of 2. Solutions are given in the following table.

$$\begin{aligned}
\bmod 2^1 &: \quad x \equiv 1 \\
\bmod 2^2 &: \quad x \equiv 1, 3 \\
\bmod 2^3 &: \quad x \equiv 1, 3, 5, 7 \\
\bmod 2^4 &: \quad x \equiv 3, 5, 11, 13 \\
\bmod 2^5 &: \quad x \equiv 5, 11, 21, 27
\end{aligned}$$

It is not hard to see that for $n \ge 3$, there are always 4 solutions mod 2^n. If $\mathbb{Z}/2^n\mathbb{Z}$ were a field, this would not be possible: in a field, a quadratic equation has at most two solutions. However, two of the four are sort of spurious: they do not "lift" to solutions mod 2^{n+1}. Now if we pass to the limit and only consider solutions that can be lifted indefinitely, then we find two solutions, as expected.

19.2. **Definition.** The ring $\mathbb{Z}_p$ of *p-adic integers* is

$$\mathbb{Z}_p = \{(a_n) : a_n \equiv a_{n+1} \bmod p^n \text{ for all } n \geq 1\} \subset \prod_{n=1}^{\infty} \mathbb{Z}/p^n\mathbb{Z}\,.$$

There is a canonical inclusion $\mathbb{Z} \hookrightarrow \mathbb{Z}_p$, given by

$$a \longmapsto (\bar{a}, \bar{a}, \bar{a}, \dots)\,.$$

Now we need some structural information on the ring $\mathbb{Z}_p$.

19.3. **Theorem.** *$\mathbb{Z}_p$ is an integral domain. It only has one maximal ideal, $p\mathbb{Z}_p$, and all non-zero ideals have the form $p^n\mathbb{Z}_p$ for some $n \geq 0$. (In particular, $\mathbb{Z}_p$ is a PID and therefore a UFD.) Its unit (or multiplicative) group is $\mathbb{Z}_p^\times = \mathbb{Z}_p \setminus pZ_p$.*

Proof. (a) $p\mathbb{Z}_p$ is a maximal ideal. We show that $\mathbb{Z}_p/p\mathbb{Z}_p \cong \mathbb{Z}/p\mathbb{Z}$; since the latter is a field, the claim follows. Consider the map

$$\mathbb{Z}_p/p\mathbb{Z}_p \ni (a_1, a_2, \dots) + p\mathbb{Z}_p \longmapsto a_1 \in \mathbb{Z}/p\mathbb{Z}\,.$$

It is a well-defined ring homomorphism and obviously surjective. The homomorphism $\mathbb{Z} \to \mathbb{Z}_p$ induces a homomorphism $\mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}_p/p\mathbb{Z}_p$, which is inverse to the map above, hence we have an isomorphism.

(b) We have of course that $\mathbb{Z}_p^\times \subset \mathbb{Z}_p \setminus p\mathbb{Z}_p$ (an element in a maximal ideal cannot be a unit). Let us show that we actually have equality. So take $u \in \mathbb{Z}_p \setminus p\mathbb{Z}_p$. If $u = (u_1, u_2, \dots)$, each u_n is invertible in $\mathbb{Z}/p^n\mathbb{Z}$, so there are unique v_n such that $u_n v_n = 1$; then $v = (v_1, v_2, \dots) \in \mathbb{Z}_p$ and $uv = 1$.

(c) We now see easily that $p\mathbb{Z}_p$ is the *only* maximal ideal. For assume that $\mathfrak{m}$ is another maximal ideal. Then $\mathfrak{m} \setminus p\mathbb{Z}_p \neq \emptyset$, and by (b), this means that $\mathfrak{m}$ contains a unit, hence $\mathfrak{m} = \mathbb{Z}_p$, a contradiction.

(d) We have $\bigcap_{n\geq 1} p^n\mathbb{Z}_p = \{0\}$. For $a = (a_1, a_2, \dots) \in p^n\mathbb{Z}_p$ implies $a_j = 0$ for $j \leq n$.

(e) If $a \in \mathbb{Z}_p \setminus \{0\}$, then there is some $n \geq 0$ and some $u \in \mathbb{Z}_p^\times$ such that $a = p^n u$: By (d), there is some n such that $a \in p^n\mathbb{Z}_p \setminus p^{n+1}\mathbb{Z}_p$. Then $a = p^n u$ where $u \in \mathbb{Z}_p \setminus p\mathbb{Z}_p = \mathbb{Z}_p^\times$.

(f) Let $I \subset \mathbb{Z}_p$ be a non-zero ideal. Then, by (d) again, there is some n such that $I \subset p^n\mathbb{Z}_p$, but $I \not\subset p^{n+1}\mathbb{Z}_p$. So there is some $a \in I$ such that $a = p^n u$ with $u \in \mathbb{Z}_p^\times$. Since u is invertible, $p^n = au^{-1} \in I$ as well, and we find $p^n\mathbb{Z}_p \subset I$, hence $I = p^n\mathbb{Z}_p$.

(g) $\mathbb{Z}_p$ is an integral domain. Suppose $ab = 0$ with $a = (a_1, a_2, \dots)$, $b = (b_1, b_2, \dots)$. Assume $a \neq 0$; then $a = p^N u$ with some $N \geq 0$, $u \in \mathbb{Z}_p^\times$. Then $ab = 0$ implies $p^N b = 0$. Now this says that $p^N b_{n+N} = 0$ in $\mathbb{Z}/p^{N+n}\mathbb{Z}$, so $b_n \equiv b_{N+n} \equiv 0 \bmod p^n$, hence $b_n = 0$, for all n. □

Part (e) in the proof motivates the following definition.

19.4. **Definition.** For $a = (a_1, a_2, \dots) \in \mathbb{Z}_p$ define the *p-adic valuation*

$$v_p(a) = \max(\{0\} \cup \{n \geq 1 : a_n = 0\}) \in \{0, 1, \dots, \infty\}\,.$$

Then $a = p^{v_p(a)}u$ with $u \in \mathbb{Z}_p^\times$, if $a \neq 0$ (and $v_p(0) = \infty$) and the valuation is compatible with the p-adic valuation on $\mathbb{Z}$.

Define the *p-adic absolute value* by

$$|0|_p = 0\,, \quad |a|_p = p^{-v_p(a)} \quad \text{if } a \neq 0.$$

19.5. **Definition.** The *field $\mathbb{Q}_p$ of p-adic numbers* is the field of fractions of $\mathbb{Z}_p$.

We have that $\mathbb{Q}_p = \mathbb{Z}_p[1/p]$, and we can extend the p-adic valuation and absolute value to $\mathbb{Q}_p$: $v_p(a/b) = v_p(a) - v_p(b)$ and $|a/b|_p = |a|_p/|b|_p$; then for all $a \in \mathbb{Q}_p^\times$,

$$a = p^{v_p(a)}u$$

with some $u \in \mathbb{Z}_p^\times$.

19.6. **Lemma.**

(1) $|ab|_p = |a|_p\,|b|_p$.
(2) $|a+b|_p \le \max\{|a|_p, |b|_p\} \le |a|_p + |b|_p$.

Proof. Easy. □

In particular, $|\cdot|_p$ defines a metric on $\mathbb{Z}_p$ and $\mathbb{Q}_p$: $d(a,b) = |a-b|_p$. It is a fact that with this metric, $\mathbb{Z}_p$ is a compact metric space, and $\mathbb{Z}$ is dense in $\mathbb{Z}_p$. Also, $\mathbb{Q}_p$ can be identified with the completion of $\mathbb{Q}$ with respect to the p-adic absolute value $|\cdot|_p$ (in the same way as $\mathbb{R}$ is the completion of $\mathbb{Q}$ with respect to the usual absolute value $|\cdot| = |\cdot|_\infty$).

19.7. **Remark.** Define $|x|_\infty = |x|$ for $x \in \mathbb{R}$. Then for all $a \in \mathbb{Q}^\times$,

$$\prod_{v=p,\infty} |a|_v = 1\,.$$

This is easy to see. Despite its apparent triviality, this *Product Formula* (and its generalization to algebraic number fields) plays an important role in number theory.

19.8. **Lemma.**

(1) *Every series $\sum_{n=0}^\infty c_n p^n$ with $c_n \in \mathbb{Z}_p$ converges in $\mathbb{Z}_p$.*
(2) *Every $a \in \mathbb{Z}_p$ can be written uniquely in the form*

$$a = \sum_{n=0}^\infty c_n p^n$$

with $c_n \in \{0, 1, \dots, p-1\}$.

Proof. Exercise. □

As an example, we have in $\mathbb{Z}_3$

$$-2 = 1 + 2\cdot 3 + 2\cdot 3^2 + 2\cdot 3^3 + \dots\,.$$

19.9. **Proposition.** *Let $F \in \mathbb{Z}[X_1, \dots, X_k]$.*

(1) $$\forall n \ge 1\, \exists (x_1, \dots, x_k) \in \mathbb{Z}^k : p^n \mid F(x_1, \dots, x_k)$$
$$\iff \exists (x_1, \dots, x_k) \in \mathbb{Z}_p^k : F(x_1, \dots, x_k) = 0\,.$$

(2) *If F is homogeneous, we have*

$$\forall n \ge 1\, \exists (x_1, \dots, x_k) \in \mathbb{Z}^k \setminus (p\mathbb{Z})^k : p^n \mid F(x_1, \dots, x_k)$$
$$\iff \exists (x_1, \dots, x_k) \in \mathbb{Z}_p^k \setminus (p\mathbb{Z}_p)^k : F(x_1, \dots, x_k) = 0$$
$$\iff \exists (x_1, \dots, x_k) \in \mathbb{Q}_p^k \setminus \{0\} : F(x_1, \dots, x_k) = 0\,.$$

Proof. To prove the nontrivial direction ("⇒"), consider the rooted tree with nodes $(n, (\bar{x}_1, \ldots, \bar{x}_k))$ (at distance n from the root $(0, (0, \ldots, 0))$) for solutions modulo p^n, where nodes at levels n and $n+1$ are connected if the solution at the upper level reduces to the solution at the lower level mod p^n (compare the motivating example at the beginning of the section, where $p = 2$, $k = 1$, and $F = X_1^2 + 7$). Then use *König's Lemma* (see below) that says that an infinite, finitely branched rooted tree has an infinite path starting at the root. This path corresponds to a k-tuple of p-adic integers. □

In order to complete this proof, we need to prove König's Lemma.

19.10. **Theorem (König's Lemma).** *Let T be an infinite, but finitely branched, rooted tree. Then T has an infinite branch (starting at the root).*

Proof. We construct an infinite branch inductively. Let $T_1, \ldots, T_m$ be the finitely many subtrees connected to the root of T. Since T is infinite, (at least) one of the T_j must be infinite. Now the first step of the branch we construct leads to the root of T_j, and we continue from there. Since T_j is again infinite, this construction will never come to an end, thus leading to an infinite branch in T. □

Note that the proof needs the Axiom of Choice, unless there is some additional structure that we can use in order to pick one of the infinite subtrees. In our application, we can represent the nodes by tuples of integers between 0 and p^n and then pick the smallest one with respect to lexicographical ordering. So we can do without the Axiom of Choice here.

19.11. **Corollary.** *$\mathbb{Z}_p$ is compact (and hence complete) in the topology induced by the metric $d(x, y) = |x - y|_p$.*

Proof. Since $\mathbb{Z}_p$ is a metric space, we can start with an open covering consisting of open balls $B_x(r) = \{y \in \mathbb{Z}_p : |x - y|_p < r\}$. Let T_0 be the rooted tree whose nodes at level n correspond to the elements of $\mathbb{Z}/p^n\mathbb{Z}$, with node a at level $n+1$ connected to node b at level n if and only if a reduces to b mod p^n. Note that an open ball $B_x(p^{-n})$ corresponds to the subtree of T_0 whose root is the node x mod p^{n+1}. Let T be the tree obtained from T_0 by removing the subtrees (except their roots) corresponding to the balls in the given open covering. An infinite branch in T would correspond to an element of $\mathbb{Z}_p$ that is not in any of the open balls; since the balls form a covering, such an infinite branch does not exist. By König's Lemma, T must then be finite, and the finitely many leaves of T correspond to a finite subcovering of the given covering. □

The following result is one of the most important ones in the theory of p-adic numbers.

19.12. **Lemma (Hensel's Lemma).** *If $f \in \mathbb{Z}[x]$ (or $\mathbb{Z}_p[x]$) and f has a simple zero a mod p, then f has a unique (simple) zero $\alpha \in \mathbb{Z}_p$ such that $\alpha \equiv a \bmod p$.*

Proof. Strangely enough, the idea of this proof comes from Newton's method for approximating roots of polynomials. In the present context, closeness is measured by the p-adic absolute value $|\cdot|_p$.

First note that if $\alpha + p\mathbb{Z}_p = a \in \mathbb{F}_p$ (for $\alpha \in \mathbb{Z}_p$), then $f'(\alpha)$ reduces to $f'(a) \neq 0$ in $\mathbb{F}_p$; therefore $f'(\alpha)$ is invertible in $\mathbb{Z}_p$, and $v_p(f'(\alpha)) = 0$. Now let $\alpha_0 \in \mathbb{Z}_p$ be any element such that its image in $\mathbb{F}_p$ is a, and define recursively

$$\alpha_{n+1} = \alpha_n - f'(\alpha_n)^{-1} f(\alpha_n)\,.$$

I claim that (α_n) converges in $\mathbb{Z}_p$. Note that

$$f(y) - f(x) = (y-x)f'(x) + (y-x)^2 g(x,y)$$

with a polynomial $g \in \mathbb{Z}_p[x,y]$, and so

$$\begin{aligned} f(\alpha_{n+1}) &= f(\alpha_n) + (\alpha_{n+1} - \alpha_n) f'(\alpha_n) + (\alpha_{n+1} - \alpha_n)^2 g(\alpha_n, \alpha_{n+1}) \\ &= f'(\alpha_n)^{-2} f(\alpha_n)^2 g(\alpha_n, \alpha_{n+1}) \end{aligned}$$

This shows that $v_p(f(\alpha_{n+1})) \geq 2v_p(f(\alpha_n))$, and since $v_p(f(\alpha_0)) \geq 1$, we have $v_p(f(\alpha_n)) \geq 2^n$. This implies that $v_p(\alpha_{n+1} - \alpha_n) \geq 2^n$, and so (by the ultrametric triangle inequality), the sequence (α_n) is a Cauchy sequence. Since $\mathbb{Z}_p$ is complete, (α_n) converges; let α be the limit. Note that $\bar{\alpha} = a$, since $v_p(\alpha - \alpha_0) \geq 1$. Also, polynomials are continuous (in the p-adic topology), so, passing to the limit in the recursion above, we obtain

$$\alpha = \alpha - f'(\alpha)^{-1} f(\alpha) \qquad \Longrightarrow \qquad f(\alpha) = 0\,.$$

To show uniqueness, assume that α and α' are two distinct zeros of f both reducing to a mod p. Then $1 \leq n = v_p(\alpha' - \alpha) < \infty$. But we have

$$0 = f(\alpha') - f(\alpha) = (\alpha' - \alpha) f'(\alpha) + (\alpha' - \alpha)^2 g(\alpha', \alpha)$$

and so

$$f'(\alpha) = -(\alpha' - \alpha)\, g(\alpha', \alpha)\,.$$

But $v_p(f'(\alpha)) = 0$, whereas the valuation of the right hand side is at least $n > 0$, a contradiction. □

Here is an easy consequence.

19.13. **Lemma.** *Let p be an odd prime and $a \in \mathbb{Z}_p$ such that $p \nmid a$. Then a is a square in $\mathbb{Z}_p$ if and only if a is a quadratic residue mod p.*

If $a \in \mathbb{Z}_2$ is odd, then a is a square in $\mathbb{Z}_2$ if and only if $a \equiv 1 \bmod 8$.

Proof. Necessity is clear in both cases. For odd p, we consider $f(x) = x^2 - a$. If a is a quadratic residue mod p, then there is some $s \in \mathbb{F}_p$ such that $f(s) = 0$; also $f'(s) = 2s \neq 0$. By Hensel's Lemma 19.12, sufficiency follows.

For $p = 2$, we consider $f(x) = 2x^2 + x - A$, where $a = 8A + 1$. Obviously, $2 \mid f(A)$ and $2 \nmid f'(A) = 4A+1$, hence again by Hensel's Lemma 19.12, f has a root $\alpha \in \mathbb{Z}_p$. But then we also have $(4\alpha + 1)^2 - a = 8f(\beta) = 0$. □

For example, this shows that -7 is a square in $\mathbb{Z}_2$.

19.14. **Lemma.** *Let $a \in \mathbb{Q}_p^\times$ and write $a = p^n u$ with $u \in \mathbb{Z}_p^\times$ (and $n = v_p(a)$). Then a is a square in $\mathbb{Q}_p$ if and only if n is even and u is a square in $\mathbb{Z}_p$.*

Proof. Sufficiency is clear. If $a = b^2$, then $n = v_p(a) = 2v_p(b)$ must be even, and $u = (b/p^{n/2})^2 \in \mathbb{Z}_p^\times$ is a square in $\mathbb{Q}_p$. But we have $v_p(b/p^{n/2}) = 0$, so u is the square of an element in $\mathbb{Z}_p$. □

We can deduce that for a ternary quadratic form $ax^2 + by^2 + cz^2$ with abc squarefree, the necessary conditions in 18.2 imply (and therefore are equivalent to the statement) that there are nontrivial solutions in $\mathbb{R}$ and in $\mathbb{Q}_p$ for all primes p. This is clear for $\mathbb{R}$. For p an odd prime, the conditions give us a solution mod p such that $p \nmid \gcd(x, y, z)$, which then lifts to a solution in $\mathbb{Z}_p$. For $p = 2$, the conditions allow us to find a solution mod 8, which then lifts to $\mathbb{Z}_2$.

19.15. **Theorem.** *Let $Q(x, y, z)$ be a non-degenerate ternary quadratic form. Then $Q(x, y, z) = 0$ has a primitive integral solution if and only if it has nontrivial solutions in real numbers and in p-adic numbers for all primes p.*

Proof. There is a diagonal ternary quadratic form $Q' = ax^2 + by^2 + cz^2$ with abc squarefree that is equivalent to Q. It is clear that Q' has nontrivial solutions in $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{Q}_p$ if and only if Q does. So Q' satisfies the conditions in 18.2. By Legendre's Theorem 18.3, Q' has a primitive integral solution, hence a nontrivial soution in $\mathbb{Q}$. Therefore, Q also has a nontrivial solution in $\mathbb{Q}$, which can be scaled to give a primitive integral solution. □

This result is called the *Hasse* or *Local-Global Principle* for ternary quadratic forms. It states that the existence of "local" solutions (in $\mathbb{R}$, $\mathbb{Q}_p$) implies the existence of "global" solutions (in $\mathbb{Q}$). In fact, this is valid for quadratic forms in general, but the proof is nontrivial for four or more variables.

Note that this implies that a quadratic form in five or more variables has a primitive integral solution if and only if it is indefinite (i.e., has nontrivial real solutions): there are always p-adic solutions for all p in this case (Exercise!).

Note also that the Hasse Principle does not hold in general. A famous counterexample (due to Selmer) is given by $3x^3 + 4y^3 + 5z^3 = 0$, which has nontrivial solutions in $\mathbb{R}$ and all $\mathbb{Q}_p$ (Exercise), but not in $\mathbb{Q}$ (hard).

20. The Hilbert Norm Residue Symbol

There is a connection between the various necessary "local" conditions for the existence of a solution for a non-degenerate ternary quadratic form. Since every non-degenerate ternary quadratic form is equivalent to one of the form

$$a\,x^2 + b\,y^2 - z^2$$

(first diagonalize, then scale the form and/or the variables appropriately), it suffices to consider such forms.

20.1. **Definition (Hilbert Norm Residue Symbol).** Let $a, b \in \mathbb{Z} \setminus \{0\}$. For a prime p, set

$$\left(\frac{a,b}{p}\right) = 1$$

if $a\,x^2 + b\,y^2 = z^2$ has a non-trivial solution in $\mathbb{Q}_p$. Otherwise, we set

$$\left(\frac{a,b}{p}\right) = -1\,.$$

Similarly,

$$\left(\frac{a,b}{\infty}\right) = 1$$

if there is a non-trivial real solution, and -1 otherwise.

Theorem 19.15 then says that $a\,x^2 + b\,y^2 = z^2$ has a non-trivial solution in $\mathbb{Q}$ if and only if

$$\left(\frac{a,b}{v}\right) = 1 \qquad \text{for all "places" } v = p, \infty.$$

20.2. **Theorem.** *Let $a, b \in \mathbb{Z} \setminus \{0\}$. Then for almost all primes p,*

$$\left(\frac{a,b}{p}\right) = 1\,,$$

and we have the Product Formula

$$\prod_{v=p,\infty} \left(\frac{a,b}{v}\right) = 1\,.$$

(Here v runs through all primes and ∞.)

This means that the number of "places" v such that the local condition fails is always *even*.

There are always solutions in $\mathbb{Q}_p$ if p is odd and does not divide ab. By definition, this implies that only finitely many of the symbols can be -1.

The product formula then is equivalent to the Law of Quadratic Reciprocity and its supplements.

20.3. **Useful Fact.** Here is a simple but useful fact: for all nonzero $a, b, c \in \mathbb{Z}$ and all "places" v, we have

$$\left(\frac{ab,ac}{v}\right) = \left(\frac{ab,-bc}{v}\right).$$

This is because the forms $ab\,x^2 + ac\,y^2 - z^2$ and $ab\,x^2 - bc\,y^2 - z^2$ are equivalent (multiply the form by $-ab$, then scale x and y to remove the squares in the coefficients, then exchange x and z).

In the following, we will assume that a and b are squarefree. This is no loss of generality, since we can achieve this by scaling the variables.

20.4. **Values of the Hilbert Norm Residue Symbol, I.** Let p be an odd prime. Assume u and v are not divisible by p. Then we have

$$\left(\frac{u,v}{p}\right) = 1\,, \quad \left(\frac{u,vp}{p}\right) = \left(\frac{u}{p}\right), \quad \left(\frac{up,v}{p}\right) = \left(\frac{v}{p}\right), \quad \left(\frac{up,vp}{p}\right) = \left(\frac{-uv}{p}\right).$$

By an easy generalization of Lemma 15.7, we get the first equality. The next two are equivalent, the symbol being obviously symmetric. We note that any non-trivial solution of $u\,x^2 + vp\,y^2 = z^2 \bmod p^2$ must have x and z not divisible by p. But then a solution mod p forces u to be a quadratic residue mod p. On the other hand, if u is a quadratic residue mod p, then there is a solution $(x,y,z) = (1,0,z)$ in $\mathbb{Z}_p$ by Lemma 19.13. For the last equality, we make use of 20.3, with $a = p$, $b = u$, $c = v$; we are then back in the previous case.

20.5. **Values of the Hilbert Norm Residue Symbol, II.** For $p = 2$, the situation is the most complicated. Let u and v be odd. Then

$$\left(\frac{u,v}{2}\right) = (-1)^{\frac{u-1}{2}\,\frac{v-1}{2}}\,, \quad \left(\frac{u,2v}{2}\right) = (-1)^{\frac{u^2-1}{8}}(-1)^{\frac{u-1}{2}\,\frac{v-1}{2}}\,,$$
$$\left(\frac{2u,v}{2}\right) = (-1)^{\frac{v^2-1}{8}}(-1)^{\frac{u-1}{2}\,\frac{v-1}{2}}\,, \quad \left(\frac{2u,2v}{2}\right) = (-1)^{\frac{u^2v^2-1}{8}}(-1)^{\frac{u-1}{2}\,\frac{v-1}{2}}\,.$$

This can be verified with the help of Lemma 19.13, in a similar spirit as before.

20.6. **Values of the Hilbert Norm Residue Symbol, III.** Finally, we have to deal with $p = \infty$. Here the situation is simple:

$$\left(\frac{a,b}{\infty}\right) = -1 \iff a < 0 \text{ and } b < 0\,.$$

20.7. **Proof of the Product Formula.** We want to prove that

$$\prod_{v=p,\infty} \left(\frac{a,b}{v}\right) = 1\,.$$

We first observe that the symbols are *bimultiplicative*:

$$\left(\frac{aa',b}{v}\right) = \left(\frac{a,b}{v}\right)\left(\frac{a',b}{v}\right) \quad \text{and} \quad \left(\frac{a,bb'}{v}\right) = \left(\frac{a,b}{v}\right)\left(\frac{a,b'}{v}\right).$$

This follows from the values given above (but there is really a deeper reason for that). It therefore suffices to prove the product formula in each of the following cases (where p and q are distinct odd primes):

$$(a,b) = (-1,-1),\ (-1,2),\ (-1,p),\ (2,2),\ (2,p),\ (p,p),\ (p,q)\,.$$

Of these, the cases $(2,2)$ and (p,p) can be reduced to $(-1,2)$ and $(-1,p)$, respectively, using 20.3 again. The remaining cases are shown in the following table. (All symbols not shown here are $+1$.)

(a,b)	$\left(\frac{a,b}{\infty}\right)$	$\left(\frac{a,b}{2}\right)$	$\left(\frac{a,b}{p}\right)$	$\left(\frac{a,b}{q}\right)$
$(-1,-1)$	-1	-1	$+1$	$+1$
$(-1,2)$	$+1$	$+1$	$+1$	$+1$
$(-1,p)$	$+1$	$(-1)^{\frac{p-1}{2}}$	$\left(\frac{-1}{p}\right)$	$+1$
$(2,p)$	$+1$	$(-1)^{\frac{p^2-1}{8}}$	$\left(\frac{2}{p}\right)$	$+1$
(p,q)	$+1$	$(-1)^{\frac{p-1}{2}\,\frac{q-1}{2}}$	$\left(\frac{q}{p}\right)$	$\left(\frac{p}{q}\right)$

From this, we see that the product formula for the Hilbert Norm Residue Symbol is equivalent to the Law of Quadratic Reciprocity, together with its two supplements. In some sense, the Product Formula is a nicer way of stating these facts, because it is just one simple statement instead of three different ones.

21. Pell's Equation and Continued Fractions

21.1. **The Equation.** Let $d > 0$ be a nonsquare integer. The equation

$$x^2 - d\,y^2 = 1 \quad \text{or} \quad x^2 - d\,y^2 = \pm 1\,,$$

to be solved in integers x and y, is known as *Pell's Equation*. The name goes back to Euler, who for some reason mistakenly assumed that Pell had contributed to the theory of its solution. In fact, it was Euler who did most of that (but already Fermat had studied it), so "Euler's Equation" might be more appropriate...

Exercise: Find all solutions when $d \leq 0$ or d is a square!

Let us look at some examples. For $d = 2$, we have the following solutions (of $x^2 - dy^2 = 1$) with $x, y \geq 0$.

x	1	3	17	99	577	3363	19601
y	0	2	12	70	408	2378	13860

For $d = 3$, we have

x	1	2	7	26	97	362	1351
y	0	1	4	15	56	209	780

And for $d = 409$, the first two solutions are

x	1	2505297727309242798604 9
y	0	123878999864721858216 0

We observe that there are nontrivial solutions, and also that they grow quite fast (the number of digits seems to grow linearly, so the numbers grow exponentially). The sequence of solutions seems to go on, so it appears that there are infinitely many solutions. However, the first nontrivial solution may be rather large.

21.2. **Some structure.** When we studied sums of two squares, we made use of the fact that

$$x^2 + y^2 = (x + iy)(x - iy) = |x + iy|^2\,.$$

Likewise, we can study the ring

$$R_d = \mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} : a, b \in \mathbb{Z}\} \subset \mathbb{R}$$

and observe that

$$x^2 - d\,y^2 = (x + y\sqrt{d})(x - y\sqrt{d})\,.$$

More generally, if R is a ring that as an additive group is a finitely generated free $\mathbb{Z}$-module (and similary for a ring that is a finite-dimensional F-algebra for some field F) and $\alpha \in R$, then left multiplication by α,

$$m_\alpha : R \ni x \longmapsto \alpha x \in R$$

defines an endomorphism of R as a $\mathbb{Z}$-module; its determinant is called the *norm* of α,

$$N(\alpha) = \det(m_\alpha)\,.$$

Since $m_{\alpha\beta} = m_\alpha \circ m_\beta$, it follows that

$$N(\alpha\beta) = \det(m_{\alpha\beta}) = \det(m_\alpha \circ m_\beta) = \det(m_\alpha)\det(m_\beta) = N(\alpha)N(\beta)\,,$$

i.e., the norm is multiplicative.

For R_d, we obtain, representing m_α by the matrix M_α with respect to the $\mathbb{Z}$-basis $1, \sqrt{d}$:

$$N(x+y\sqrt{d}) = \begin{vmatrix} x & y \\ dy & x \end{vmatrix} = x^2 - d\,y^2\,.$$

Since (by the formula for the inverse of a 2×2 matrix)

$$M^{-1}_{x+y\sqrt{d}} = N(x+y\sqrt{d})^{-1} M_{x-y\sqrt{d}}\,,$$

we see that $N(x+y\sqrt{d}) = \pm 1$ if and only if $x+y\sqrt{d} \in R_d^\times$ is a unit. This means that what we are interested in is essentially the unit group of the ring R_d. In particular, we see that the solution sets

$$S_d = \{(x,y) \in \mathbb{Z}^2 : x^2 - dy^2 = 1\} \quad \text{and} \quad T_d = \{(x,y) \in \mathbb{Z}^2 : x^2 - dy^2 = \pm 1\}$$

have a natural structure as abelian groups under the "multiplication"

$$(x,y) * (x',y') = (xx' + dyy', xy' + yx')\,.$$

Also, S_d is a subgroup of T_d of index at most 2.

Let

$$\phi : S_d \longrightarrow \mathbb{R}^\times\,, \quad (x,y) \longmapsto x + y\sqrt{d}\,.$$

Then $1/\phi(x,y) = x - y\sqrt{d}$ and therefore

$$x = \frac{\phi(x,y) + 1/\phi(x,y)}{2}\,, \qquad y = \frac{\phi(x,y) - 1/\phi(x,y)}{2\sqrt{d}}\,.$$

This shows that ϕ is injective and that $\phi(x,y) > 0$ if and only if $x > 0$, and $\phi(x,y) > 1$ if and only if $x, y > 0$. Since $(-1,0) \in S_d$, the homomorphism $S_d \to \{\pm 1\}$, $(x,y) \mapsto \operatorname{sign}(x+y\sqrt{d})$ is onto, and the subgroup

$$S_d^+ = \{(x,y) \in S_d : x > 0\}$$

is of index 2 in S_d.

21.3. **Lemma.** *Assume that S_d^+ is nontrivial and let (x_1, y_1) be the solution with minimal x_1 such that $x_1, y_1 > 0$. Then (x_1, y_1) generates S_d^+.*

Proof. Note that $\alpha = \phi(x_1, y_1) > 1$ and that there is no $(x,y) \in S_d^+$ such that $1 < \phi(x,y) < \alpha$ (since $0 < x' < x$, $0 \le y, y'$ implies $\phi(x,y) < \phi(x',y')$). Now let $(x,y) \in S_d^+$ be arbitrary and set $\beta = \phi(x,y)$. Since $\alpha > 1$, there is $n \in \mathbb{Z}$ such that $\alpha^n \le \beta < \alpha^{n+1}$. This implies that

$$1 \le \alpha^{-n}\beta = \phi((x,y) * (x_1,y_1)^{-n}) < \alpha$$

and so $(x,y) * (x_1,y_1)^{-n} = (1,0)$, hence $(x,y) = (x_1,y_1)^n$. □

The idea of this proof is that a discrete subgroup of $(\mathbb{R}, +)$ (take the logarithm to get into the additive group of $\mathbb{R}$) is either trivial or cyclic.

We see that *if* there are nontrivial solutions, then $S_d \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}$ as an abstract abelian group. It remains to show that there are nontrivial solutions.

21.4. **Diophantine approximation.** Note that a solution to $x^2 - d\,y^2 = 1$ will provide a very good rational approximation of $\sqrt{d}$:

$$\left|\sqrt{d} - \frac{x}{y}\right| = \frac{|x^2 - dy^2|}{y^2|\sqrt{d} + x/y|} < \frac{1}{2\sqrt{d}y^2}$$

So the question is whether such good approximations always exist. That we can get at least close is shown by the following result.

Lemma. *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational real number. Then there are infinitely many rational numbers p/q such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}\,.$$

Proof. We make use of the "box principle". Let $\langle x\rangle = x - \lfloor x\rfloor$ denote the fractional part of $x \in \mathbb{R}$. Then of the $n+1$ numbers

$$0, \langle\alpha\rangle, \langle 2\alpha\rangle, \ldots, \langle n\alpha\rangle$$

in the half-open interval $[0,1)$, there must be two that fall into the same subinterval $[k/n, (k+1)/n)$ for some $0 \le k < n$. So there are $0 \le l < m \le n$ such that

$$\frac{1}{n} > |\langle m\alpha\rangle - \langle l\alpha\rangle| = |(m-l)\alpha - (\lfloor m\alpha\rfloor - \lfloor l\alpha\rfloor)|\,.$$

Setting $p = \lfloor m\alpha\rfloor - \lfloor l\alpha\rfloor$ and $q = m - l$, we find

$$0 < \left|\alpha - \frac{p}{q}\right| < \frac{1}{nq} \le \frac{1}{q^2}\,.$$

(Note that α is irrational, so not equal to p/q.) By taking n larger and larger, we find a sequence of fractions p/q such that $|\alpha - p/q| < 1/q^2$ and $q \to \infty$ (since $|\alpha - p/q| \to 0$). □

21.5. **Remarks.**

(1) The property that the set

$$\left\{\frac{p}{q} \in \mathbb{Q} : \left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}\right\}$$

is infinite in fact characterizes the *irrational* numbers among the real numbers α. In fact, if $\alpha = r/s$, then $|\alpha - p/q|$ is either zero or else at least $1/qs$, so $0 < |\alpha - p/q| < 1/q^2$ implies $q < s$.

(2) If $\alpha \in \mathbb{R}$ is *algebraic* (i.e., α is a root of a monic polynomial with rational coefficients), then α cannot approximated much better by rational numbers than in the result above. More precisely, for every $\varepsilon > 0$, there are only finitely many fractions p/q such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\varepsilon}}\,.$$

This is known as *Roth's Theorem* (or also the *Thue-Siegel-Roth Theorem*, since Thue and Siegel obtained similar, but weaker results earlier) and a quite deep result. It implies for example that an equation

$$F(x,y) = m$$

where $F \in \mathbb{Z}[x,y]$ is homogeneous of degree ≥ 3 and $0 \ne m \in \mathbb{Z}$ can have only finitely many integral solutions. (Such equations are called *Thue Equations*; Thue used his result mentioned above to prove this statement.)

21.6. **Application to Pell's Equation.** In order to show that nontrivial solutions to Pell's Equation exist, we use the result we just prove as a starting point and apply the box principle two more times. First let us show the following.

There are infinitely many pairs $(x, y) \in \mathbb{Z}^2$ *such that* $|x^2 - dy^2| < 2\sqrt{d} + 1$.

Indeed, by Lemma 21.4 there are infinitely many (x, y) such that $|x/y - \sqrt{d}| < 1/y^2$ (note that $\sqrt{d}$ is irrational, since d is not a square). Then

$$|x^2 - d\,y^2| = y^2 \Big|\frac{x}{y} - \sqrt{d}\Big|\Big(\frac{x}{y} + \sqrt{d}\Big) < 2\sqrt{d} + \Big|\frac{x}{y} - \sqrt{d}\Big| < 2\sqrt{d} + \frac{1}{y^2} \le 2\sqrt{d} + 1\,.$$

Now there are only finitely many integers m such that $|m| < 2\sqrt{d} + 1$. Therefore there must be one m such that there are infinitely many pairs (x, y) such that $x^2 - dy^2 = m$ (note the use of the box principle).

Now the idea is to "divide" two suitably chosen solutions to $x^2 - dy^2 = m$ to obtain a solution of $x^2 - dy^2 = 1$. To this end, we use the box principle another time to conclude that there must be two pairs (x, y) and (u, v) with $0 < x < u$, $0 < y$, $0 < v$, $x^2 - dy^2 = u^2 - dv^2 = m$ and $x \equiv u$, $y \equiv v \bmod m$. Then

$$(xu - d\,yv)^2 - d(uy - xv)^2 = m^2$$

and

$$xu - d\,yv \equiv x^2 - d\,y^2 = m \equiv 0 \bmod m\,, \qquad uy - xv \equiv xy - xy = 0 \bmod m\,,$$

so

$$\Big(\frac{|xu - d\,yv|}{m}, \frac{uy - xv}{m}\Big) \in S_d^+$$

is a nontrivial solution (otherwise $u/v = x/y$ which implies $(u, v) = (x, y)$, since $y, v > 0$ and $x \perp y$, $u \perp v$). We have proved:

21.7. **Theorem.** *Pell's Equation has nontrivial solutions. The solution set* S_d *has a natural structure as an abelian group, and as such it is isomorphic to* $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}$.

The generator (x_1, y_1) of S_d^+ with $x_1, y_1 > 0$ is called the *fundamental solution* of the equation. All other solutions then are of the form

$$\pm(x_n, y_n) \quad \text{with} \quad x_n + y_n\sqrt{d} = (x_1 + y_1\sqrt{d})^n\,, \quad n \in \mathbb{Z}\,.$$

Note that for n large,

$$x_n \approx \frac{(x_1 + y_1\sqrt{d})^n}{2}\,, \qquad y_n \approx \frac{(x_1 + y_1\sqrt{d})^n}{2\sqrt{d}}\,,$$

which explains the observed linear growth in the number of digits in the solutions.

21.8. **Continued fractions.** The question remains how to actually *find* the fundamental solution to a given Pell Equation, or more generally, how to find the good rational approximations to irrational numbers that we are promised in Lemma 21.4. The answer is provided by *continued fractions.*

Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational real number again. We set $\alpha_0 = \alpha$ and then define recursively for $n \ge 0$

$$a_n = \lfloor \alpha_n \rfloor\,, \qquad \alpha_{n+1} = \frac{1}{\alpha_n - a_n}\,.$$

Note that all α_n are irrational, so we can never have $\alpha_n = a_n$. Note also that $\alpha_n > 1$ and therefore $a_n \ge 1$ as soon as $n \ge 1$.

We then have

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots + \cfrac{1}{a_{n-1} + \cfrac{1}{\alpha_n}}}}}$$

and we denote the nested fraction on the right hand side by

$$[a_0; a_1, a_2, \ldots, a_{n-1}, \alpha_n]$$

(where $a_0, a_1, \ldots, a_{n-1}, \alpha_n$ can be arbitrary numbers). Note that we have the recurrence

$$[a_0] = a_0\,, \qquad [a_0; a_1, \ldots, a_{n-2}, a_{n-1}, x] = [a_0; a_1, \ldots, a_{n-2}, a_{n-1} + 1/x] \quad (n \geq 1)\,.$$

We call the formal expression

$$[a_0; a_1, a_2, a_3, \ldots]$$

the *continued fraction expansion* of α.

Given integers $a_0, a_1, a_2, \ldots$ (with $a_1, a_2, \cdots \geq 1$), it is clear that $[a_0; a_1, \ldots, a_n]$ is a rational number. Let us find a way to compute its numerator and denominator efficiently.

21.9. **Lemma.** *Set $p_{-2} = 0$, $q_{-2} = 1$, $p_{-1} = 1$, $q_{-1} = 0$ and define recursively*

$$p_{n+1} = a_{n+1}p_n + p_{n-1}\,, \qquad q_{n+1} = a_{n+1}q_n + q_{n-1}\,.$$

Then we have the following.

(1) *$p_{n+1}q_n - p_nq_{n+1} = (-1)^n$ for all $n \geq -2$. In particular, $p_n \perp q_n$.*

(2) *$[a_0; a_1, \ldots, a_n] = p_n/q_n$ for all $n \geq 0$.*

(3) *If $[a_0; a_1, a_2, \ldots]$ is the continued fraction expansion of α, then for $n \geq 0$,*

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_nq_{n+1}} \leq \frac{1}{q_n^2}\,.$$

Also, $\operatorname{sign}(\alpha - p_n/q_n) = (-1)^n$.

(4) *Under the assumptions of (3), we have*

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \frac{p_4}{q_4} < \cdots < \alpha < \cdots < \frac{p_5}{q_5} < \frac{p_3}{q_3} < \frac{p_1}{q_1}\,.$$

The fractions p_n/q_n are called the *convergents* of the continued fraction expansion $[a_0; a_1, a_2, \ldots]$.

Proof.

(1) This is an easy induction.

(2) We claim that more generally, for $n \geq -1$ we have

$$[a_0; a_1, \ldots, a_n, x] = \frac{p_nx + p_{n-1}}{q_nx + q_{n-1}}\,.$$

This is clear for $n = -1$: $x = (p_{-1}x + p_{-1})/(q_{-1}x + q_{-2})$. Assuming it is OK for n, we find

$$\begin{aligned}[a_0; a_1, \dots, a_n, a_{n+1}, x] &= [a_0; a_1, \dots, a_n, a_{n+1} + 1/x] \\ &= \frac{p_n\big(a_{n+1} + \frac{1}{x}\big) + p_{n-1}}{q_n\big(a_{n+1} + \frac{1}{x}\big) + q_{n-1}} = \frac{(a_{n+1}p_n + p_{n-1})x + p_n}{(a_{n+1}q_n + q_{n-1})x + q_n} \\ &= \frac{p_{n+1}x + p_n}{q_{n+1}x + q_n}\end{aligned}$$

Specializing x to a_{n+1}, the statement of the lemma follows.

(3) We have $\alpha = [a_0; a_1, \dots, a_n, \alpha_{n+1}]$. Using the claim established above, we obtain

$$\alpha - \frac{p_n}{q_n} = \frac{\alpha_{n+1}p_n + p_{n-1}}{\alpha_{n+1}q_n + q_{n-1}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n(\alpha_{n+1}q_n + q_{n-1})}\,.$$

This proves the assertion about the sign. Also, $\alpha_{n+1} > a_{n+1}$, hence $\alpha_{n+1}q_n + q_{n-1} > a_{n+1}q_n + q_{n-1} = q_{n+1}$, so

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}}\,.$$

(4) Note that

$$\frac{p_{n+2}}{q_{n+2}} - \frac{p_n}{q_n} = \frac{a_{n+2}(p_{n+1}q_n - p_n q_{n+1})}{q_n q_{n+2}} = \frac{a_{n+2}(-1)^n}{q_n q_{n+2}}$$

is positive for n even and negative for n odd.

□

21.10. **Corollary.** *Let $a_0, a_1, a_2, \dots \in \mathbb{Z}$ with $a_1, a_2, \dots \geq 1$. Then the sequence $(p_n/q_n)_n$, where*

$$\frac{p_n}{q_n} = [a_0; a_1, a_2, \dots, a_n]\,,$$

converges to a limit α, and $[a_0; a_1, a_2, \dots]$ is the continued fraction expansion of α.

Proof. By the above, we have $|p_{n+1}/q_{n+1} - p_n/q_n| = 1/(q_n q_{n+1})$. Also, $q_n \geq n - 1$ for $n \geq 2$, hence $\sum_{n\geq 0} 1/(q_n q_{n+1})$ converges. This implies that the sequence is Cauchy and therefore has a limit α. Since $a_0 \leq p_n/q_n < a_0 + 1$ for $n \geq 3$, we must have $a_0 = \lfloor \alpha \rfloor$. Then

$$\alpha_1 = \frac{1}{\alpha - a_0} = \lim_{n\to\infty} [a_1; a_2, \dots, a_n]\,.$$

Continuing, we find successively that the a_n are the numbers making up the continued fraction expansion of α. □

In order to conclude that we can use the continued fraction expansion of $\sqrt{d}$ in order to find the fundamental solution to a Pell Equation, we need to know that any sufficiently good approximation appears as a convergent.

21.11. **Lemma.** *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, and let $p/q \in \mathbb{Q}$ such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}\,.$$

Then p/q is a convergent of the continued fraction expansion of α.

Proof. We first consider the case $q = 1$. Then $p/q = p$ is the integer closest to α, so $p = a_0 = p_0/q_0$ (and we are done) or $p = a_0+1$. In the latter case, $a_0+1/2 < \alpha$, and we find $\alpha_1 < 2$, so $a_1 = 1$ and $p_1/q_1 = a_0 + 1 = p$, so we are done again.

For $q \geq 2$, we claim a slightly stronger result, namely that $|\alpha-p/q| < 1/(q(2q-1))$ implies that $p/q = p_n/q_n$ for some n. We can assume that p/q is in lowest terms.

We will prove this by induction on q. First observe that we can assume that $0 < \alpha < 1$ (since shifting everything by a_0 does not affect the assumption nor the conclusion). Then we must have $0 < p/q < 1$ as well: first we cannot have equality on either end since $q > 1$. Second, assume that $p/q < 0$. Then

$$\frac{1}{q(2q-1)} > \left|\alpha - \frac{p}{q}\right| > \left|\frac{p}{q}\right| \geq \frac{1}{q}\,,$$

which contradicts $q \geq 2$. Similarly if $p/q > 1$. So we have $0 < p < q$. We now observe that

$$\left|\frac{1}{\alpha} - \frac{q}{p}\right| = \left|\alpha - \frac{p}{q}\right| \frac{q}{p\alpha} < \frac{1}{p\alpha(2q-1)}\,.$$

Also,

$$\alpha(2q-1) \geq \Big(\frac{p}{q} - \frac{1}{q(2q-1)}\Big)(2q-1) = 2p - \frac{p}{q} - \frac{1}{q} \geq 2p - 1\,,$$

so that

$$\left|\frac{1}{\alpha} - \frac{q}{p}\right| < \frac{1}{p(2p-1)}\,.$$

If $p \geq 2$, we are therefore done by induction. If $p = 1$, then we have

$$\frac{2q-2}{q(2q-1)} = \frac{1}{q} - \frac{1}{q(2q-1)} < \alpha < \frac{1}{q} + \frac{1}{q(2q-1)} = \frac{2}{2q-1}$$

and so

$$q - \tfrac{1}{2} < \frac{1}{\alpha} < q + \frac{q}{2(q-1)} < q + 1\,.$$

If $a_1 = q$, then $p_1/q_1 = 1/q$ and we are done. Otherwise, $a_1 = q - 1$ and $a_2 = 1$ and then $p_2/q_2 = 1/q$ and we are done again. □

21.12. **Theorem.** *Let $d \geq 1$ be a nonsquare integer. Then the fundamental solution to the Pell Equation*

$$x^2 - d\,y^2 = 1$$

is given by $(x_1, y_1) = (p_n, q_n)$, where $n \geq 0$ is minimal such that $p_n^2 - d\,q_n^2 = 1$. Here, p_n/q_n are the convergents of the continued fraction expansion of $\sqrt{d}$.

Proof. We have seen that any nontrivial positive solution (x, y) satisfies

$$\left|\sqrt{d} - \frac{x}{y}\right| < \frac{1}{2\sqrt{d}y^2} \leq \frac{1}{2y^2}\,.$$

By the previous lemma, every such solution must be of the form $(x, y) = (p_n, q_n)$, where p_n/q_n is a convergent of the continued fraction expansion of $\sqrt{d}$. Since all a_n are positive (including a_0), the sequence of numerators $p_0, p_1, \ldots$ is strictly

increasing. Therefore the fundamental solution (which is the smallest positive solution) must be the first one we encounter. □

21.13. **Example.** Let us illustrate the result by an example. Take $d = 31$, then $5 < \sqrt{31} < 6$. We obtain the following table.

n	α_n	a_n	p_n	q_n	$p_n^2 - 31\,q_n^2$
0	$\sqrt{31}$	5	5	1	-6
1	$\frac{1}{\sqrt{31}-5} = \frac{\sqrt{31}+5}{6}$	1	6	1	5
2	$\frac{6}{\sqrt{31}-1} = \frac{\sqrt{31}+1}{5}$	1	11	2	-3
3	$\frac{5}{\sqrt{31}-4} = \frac{\sqrt{31}+4}{3}$	3	39	7	2
4	$\frac{3}{\sqrt{31}-5} = \frac{\sqrt{31}+5}{2}$	5	206	37	-3
5	$\frac{2}{\sqrt{31}-5} = \frac{\sqrt{31}+5}{3}$	3	657	118	5
6	$\frac{3}{\sqrt{31}-4} = \frac{\sqrt{31}+4}{5}$	1	863	155	-6
7	$\frac{5}{\sqrt{31}-1} = \frac{\sqrt{31}+1}{6}$	1	1520	273	1
8	$\frac{6}{\sqrt{31}-5} = \sqrt{31}+5$				

We find the fundamental solution $(1520, 273)$. Note that if we had first found a solution to $x^2 - d\,y^2 = -1$, we can simply "square" it to find the fundamental solution to $x^2 - d\,y^2 = +1$.

Note also that it can be shown (Exercise!) that $p_n^2 - d\,q_n^2 = \pm 1$ if and only if $\alpha_{n+1} = \sqrt{d} + a$ with an integer a (which must be $\lfloor \sqrt{d} \rfloor$ unless $n = -1$).

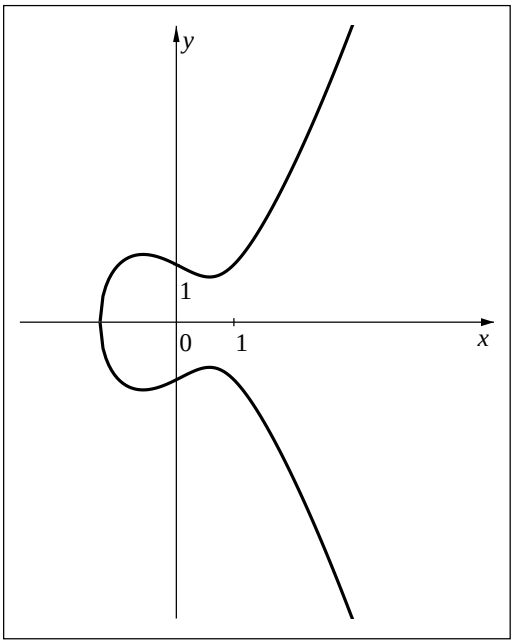
22. Elliptic Curves

After we have looked at equations of degree 2 in some detail, let us now move up one step and consider equations of degree 3. We will restrict our attention to the case of two variables. It turns out that a very interesting class of such equations can be brought into the following form.

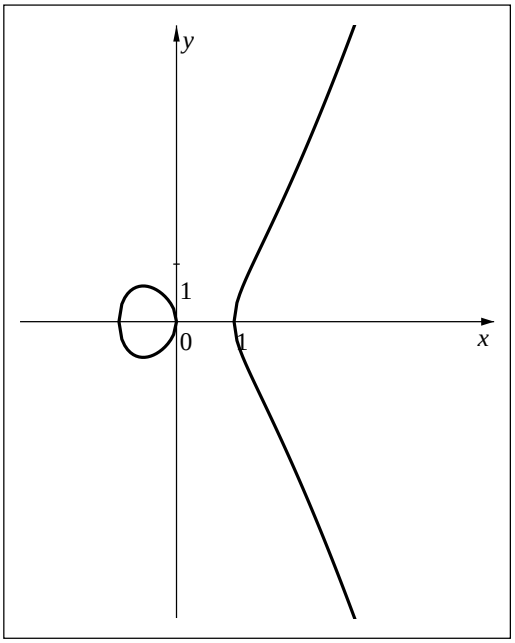
$$y^2 = x^3 + A\,x + B\,,$$

where A and B are rational numbers (which we can even assume to be integers, by scaling the variables x and y appropriately). We will see that we should require that $4A^3 + 27B^2 \neq 0$. In this case, the equation above is said to define an *elliptic curve* E.

As this is a Number Theory course, we will be interested in the *rational* solutions of the equation (also called *rational points* on E). However, it makes sense to study solutions (or points) with coordinates in any field. For example, we can easily visualize the set of real points as a curve in the xy-plane.



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



$y^2 = x^3 - x$

It turns out that it is advantageous to "complete" (or "close up", or "compactify") the curve by not considering it in the usual *affine plane*, but by considering it in the *projective plane*, which is obtained by adding a "point at infinity" to the affine plane for each direction (represented by a family of parallel lines). For a formal definition, see below.

What it comes down to is the following. We replace x and y in the equation by x/z and y/z, respectively, and then multiply by an appropriate power of z to obtain a homogeneous polynomial. For our elliptic curve, we obtain

$$y^2 = x^3 + A\,x + B \longrightarrow \left(\frac{y}{z}\right)^2 = \left(\frac{x}{z}\right)^3 + A\,\frac{x}{z} + B \longrightarrow y^2 z = x^3 + A\,xz^2 + B\,z^3\,.$$

The points are now given by *triples* (ξ, η, ζ), but we have to identify triples that are related by scaling with a constant factor (and $(0, 0, 0)$ is not allowed). We write $(\xi : \eta : \zeta)$ for the point given by (ξ, η, ζ); then we have

$$(\xi : \eta : \zeta) = (\lambda\xi : \lambda\eta : \lambda\zeta)$$

for all $(\xi, \eta, \zeta) \neq (0, 0, 0)$ and $\lambda \neq 0$. We find the points in the affine xy-plane as $(\xi : \eta : 1)$; all points with $\zeta \neq 0$ come up in that way. The remaining points (with $\zeta = 0$) are the "points at infinity"; the point $(\xi : \eta : 0)$ corresponds to the direction of lines parallel to $\eta x = \xi y$.

The set of all these points with ξ, η, ζ rational is denoted $\mathbb{P}^2(\mathbb{Q})$, where $\mathbb{P}^2$ stands for the projective plane. We then set

$$E(\mathbb{Q}) = \{(\xi : \eta : \zeta) \in \mathbb{P}^2(\mathbb{Q}) : \eta^2\zeta = \xi^3 + A\xi\zeta^2 + B\zeta^3\}\,;$$

this is the set of *rational points* on the elliptic curve E. Note that this set is always non-empty: it contains the point

$$O = (0 : 1 : 0)\,,$$

which is the only point at infinity on E. (To see this, put $z = 0$ in the homogeneous equation for E; we find $x^3 = 0$, so the points at infinity are all of the form $(0 : \eta : 0)$. Because of the identification modulo scaling, this is always the same point $(0 : 1 : 0)$.)

22.1. **Group Structure.**
We now want to define a group structure on the points on E, and in particular on $E(\mathbb{Q})$. For this, we note that *a line intersects the curve E in exactly three points, counting multiplicity.* There are some remarks to make here. First of all, "counting multiplicity" means that a point on E that has the line as its tangent has to be counted twice (or even three times, when the point is a point of inflection). Second, some of the points may have coordinates in a larger field (for example, a line may intersect $E(\mathbb{Q})$ only in one point, even counting multiplicity; the other two points will then have complex coordinates). Third, some (or all) of the intersection points may be at infinity. For example, the "line at infinity" $z = 0$ intersects E just in the point $O = (0 : 1 : 0)$, which is a point of inflection with tangent $z = 0$, and so must be counted three times. (To see that O is an inflection point, consider the affine plane on which $y \neq 0$. In terms of the curve equation, this comes down to setting y equal to 1:

$$z = x^3 + A\,xz^2 + B\,z^3\,;$$

O then has coordinates $(x, z) = (0, 0)$, and from the equation, we see that the curve looks very much like $z = x^3$ near the origin.)

But note that when two of the intersection points are rational (or real), then so must be the third: its x-coordinate (say) is a root of a cubic polynomial with rational (real) coefficients, and the other two roots are rational (real) numbers.

We now define addition on E by saying that O is the origin (zero element) of the group and that three points P_1, P_2, P_3 add up to zero if and only if they are the three points of intersection of E with a line.

Assuming this really defines a group law, how do we find the negative $-P$ of a point P, and how do we find the sum $P + Q$ of two points?

For the negative, note that $O + P + (-P) = O$, so $O, P, -P$ must be the three points of intersection of E with a line. This line is determined by the two points O and P (in the affine picture, it is the vertical line through P), and $-P$ is the third point of intersection. In terms of coordinates, if $P = (\xi, \eta)$, then $-P = (\xi, -\eta)$. In particular, $P = -P$, and so $2P = O$, if and only if $\eta = 0$ (or $P = O$).

To find the sum, note that $P + Q + (-(P + Q)) = O$, so $-(P + Q)$ must be the third point of intersection of E with the line through P and Q. (This line is the tangent line to E at P when $P = Q$.) We then find $P + Q$ as the negative of this point.

It is pretty obvious that the operation defined in this way satisfies all axioms of an abelian group except associativity. The proof of associativity is a bit involved, so we skip it here.

22.2. **Example.** Let us consider a specific curve,

$$E : y^2 = x^3 - x + 1\,.$$

(See the left hand picture on page 51.) There are some obvious points:

$$\pm P = (1, \pm 1)\,, \quad \pm Q = (0, \pm 1)\,, \quad \pm R = (-1, \pm 1)$$

Let us do some computations with them. It is clear that P, Q, R are the points of intersection of E with the line $y = 1$, so we have

$$P + Q + R = O\,.$$

Let us find $P - R = P + (-R)$. For this, we first have to find the line through $P = (1, 1)$ and $-R = (-1, -1)$; this line is $y = x$. We use the equation of the line to eliminate y form the equation of E:

$$x^2 = x^3 - x + 1 \iff x^3 - x^2 - x + 1 = (x-1)^2(x+1) = 0\,.$$

We see that the three points of intersection have x-coordinates 1 (counted twice) and -1, so (using $y = x$) the points are P (twice) and $-R$. This means that the third point of intersection is P, and

$$P - R = -P \qquad \text{and therefore} \qquad R = 2P\,.$$

Since $P + Q + R = 0$, we also find that

$$Q = -3P\,.$$

So all the points we listed at the beginning are multiples of P. Let us find some more multiples. We can compute $4P$ by doubling $2P$. So we have to find the tangent line to E at $2P = R = (-1, 1)$. What is its slope? We can find it by implicit differentiation:

$$y^2 = x^3 - x + 1 \quad \Longrightarrow \quad 2y\,dy = (3x^2 - 1)\,dx \quad \Longrightarrow \quad \frac{dy}{dx} = \frac{3x^2 - 1}{2y}\,.$$

So the slope of the tangent at $2P$ is $(3 - 1)/2 = 1$, and the line has equation $y = x + 2$. We find

$$(x+2)^2 = x^3 - x + 1 \iff x^3 - x^2 - 5x - 3 = (x+1)^2(x-3) = 0\,,$$

so the third point of intersection is $-4P = (3, 5)$, and $4P = (3, -5)$.

Note that when the line has slope λ, so its equation is $y = \lambda x + \mu$ for some μ, the cubic polynomial in x we get is of the form

$$x^3 - \lambda^2\, x^2 + \text{lower order terms}\,.$$

So λ^2 is the sum of the three roots, and we can find the x-coordinate ξ_3 of $P + Q$, where $P = (\xi_1, \eta_1)$, $Q = (\xi_2, \eta_2)$, as

$$\xi_3 = \lambda^2 - \xi_1 - \xi_2 \quad \text{and then the } y\text{-coordinate is} \quad \eta_3 = -(\lambda\xi_3 + \mu)\,.$$

Let us compute $6P$ in order to see that we can also get non-integral points. The slope of the tangent at $3P = -Q = (0, -1)$ is $-1/-2 = 1/2$, so the line is $y = \frac{1}{2}x - 1$. Hence the x-coordinate of $6P$ is $(\frac{1}{2})^2 - 0 - 0 = \frac{1}{4}$, and the y-coordinate is $-(\frac{1}{2} \cdot \frac{1}{4} - 1) = \frac{7}{8}$.

For this curve E, it can be shown that the group $E(\mathbb{Q})$ is infinite cyclic and generated by P.

Let us now put these things on a more formal basis. In the following, K will be an arbitrary field (sometimes assumed not to have characteristic 2 or 3).

22.3. **Definition.**

(1) The *affine plane* over K is the set

$$\mathbb{A}^2(K) = K^2 = \{(\xi, \eta) : \xi, \eta \in K\}\,.$$

(2) The *projective plane* over K is the set of equivalence classes

$$\mathbb{P}^2(K) = (K^3 \setminus \{(0,0,0)\})/\sim\,,$$

where the equivalence relation $\sim$ is given by

$$(\xi, \eta, \zeta) \sim (\lambda\xi, \lambda\eta, \lambda\zeta) \qquad \text{for } \lambda \in K^\times.$$

We write $(\xi : \eta : \zeta)$ for the equivalence class of (ξ, η, ζ) (and call it a *K-rational point* on the projective plane).

Note that we have an injection $\mathbb{A}^2(K) \to \mathbb{P}^2(K)$, given by $(\xi, \eta) \mapsto (\xi : \eta : 1)$; its image is the set of points $(\xi : \eta : \zeta)$ with $\zeta \neq 0$. (On that set, the inverse map is $(\xi : \eta : \zeta) \mapsto (\xi/\zeta, \eta/\zeta)$.) Points with $\zeta = 0$ are called *points at infinity*.

22.4. **Definition.**

(1) An *affine (plane algebraic) curve* over K is given by a nonzero polynomial $F \in K[x, y]$.

(2) A *projective (plane algebraic) curve* C over K is given by a nonzero homogeneous polynomial $F \in K[x, y, z]$. If F has degree d, then the curve is said to be of degree d as well. A curve of degree 1 is called a *line*. We write

$$C : F(x, y, z) = 0\,.$$

The set

$$C(K) = \{(\xi : \eta : \zeta) \in \mathbb{P}^2(K) : F(\xi, \eta, \zeta) = 0\}$$

is called the set of *K-rational points on C*.

(3) If $C : F(x, y) = 0$ is an affine curve, and the (total) degree of F is d, then

$$\tilde{C} : \tilde{F}(x, y, z) = 0 \qquad \text{where} \quad \tilde{F}(x, y, z) = z^d F\Big(\frac{x}{z}, \frac{y}{z}\Big)$$

is a projective curve of degree d, called the *projective closure* of C. (Note that $F(x, y) = \tilde{F}(x, y, 1)$.)

We will only consider projective curves in the following, but it is often convenient to give an affine equation. The curve under consideration will then be the projecetive closure of this affine curve.

22.5. **Definition.**

(1) Let $C : F(x, y, z) = 0$ be a projective curve over K, $P \in C(K)$. The point $P = (\xi : \eta : \zeta)$ is said to be a *singular point* of C if

$$\frac{\partial F}{\partial x}(\xi, \eta, \zeta) = \frac{\partial F}{\partial y}(\xi, \eta, \zeta) = \frac{\partial F}{\partial z}(\xi, \eta, \zeta) = 0\,.$$

(2) The curve C is said to be *singular*, if there is a field $K' \supset K$ (which can be taken to be an algebraic extension of K) and a singular point $P \in C(K')$. Otherwise, C is said to be *regular* or *smooth*.

The motivation behind this definition is that if not all the partial derivatives vanish, then there is a well-defined tangent line to the curve at the point P, given by

$$\frac{\partial F}{\partial x}(\xi, \eta, \zeta)\, x + \frac{\partial F}{\partial y}(\xi, \eta, \zeta)\, y + \frac{\partial F}{\partial z}(\xi, \eta, \zeta)\, z = 0\,.$$

Such a point is "nice", whereas a point without a tangent line is "bad".

Now we can define what an elliptic curve is.

22.6. **Definition.** An *elliptic curve* over K is a smooth projective curve over K given by an equation of the form

$$E : y^2z + a_1\,xyz + a_3\,yz^2 = x^3 + a_2\,x^2z + a_4\,xz^2 + a_6\,z^3$$

(with $a_1, a_2, a_3, a_4, a_6 \in K$). When the characteristic of K is not 2 or 3, then by a suitable linear change of variables, we can put this into the form

$$E : y^2z = x^3 + A\,xz^2 + B\,z^3\,.$$

(First complete the square on the left (need to divide by 2), then the cube on the right (need to divide by 3).)

We will usually just write the affine form of these equations:

$$y^2 + a_1\,xy + a_3\,y = x^3 + a_2\,x^2 + a_4\,x + a_6\,, \qquad y^2 = x^3 + A\,x + B\,.$$

A perhaps better definition is "a smooth projective curve of degree 3 with a specified K-rational point on it." It turns out that there is always a (in general nonlinear) transformation that puts this more general cubic equation into one of the form above, such that the specified point is the point at infinity $O = (0 : 1 : 0)$.

For the definition to be useful, we need a way to find out whether a given equation defines a smooth curve.

22.7. **Proposition.** *Assume that K is not of characteristic 2 or 3. A curve*

$$C : y^2z = x^3 + A\,xz^2 + B\,z^3$$

is smooth if and only if $4A^3 + 27B^2 \neq 0$.

Proof. First assume that $P = (\xi : \eta : \zeta)$ is a singular point. Then $\zeta \neq 0$ (since the only point at infinity on C is $(0 : 1 : 0)$, which is not singular), so we can take $\zeta = 1$. Let $F(x, y, z) = -y^2z + x^3 + Axz^2 + Bz^3$. Then the conditions are

$$\begin{aligned}
\frac{\partial F}{\partial x}(\xi, \eta, 1) &= 3\xi^2 + A = 0 \\
\frac{\partial F}{\partial y}(\xi, \eta, 1) &= -2\eta = 0 \\
\frac{\partial F}{\partial z}(\xi, \eta, 1) &= -\eta^2 + 2A\xi + 3B = 0
\end{aligned}$$

Since $2 \neq 0$ in K, this implies $\eta = 0$ and then $2A\xi + 3B = 3\xi^2 + A = 0$. Multiplying the second equation by $4A^2$ and plugging in the first, we find that $4A^3 + 27B^2 = 0$.

Now assume that $4A^3 + 27B^2 = 0$. If $A \neq 0$, one checks that $P = (-3B : 0 : 2A)$ is a singular point on C. If $A = 0$, then $B = 0$ as well (since $3 \neq 0$ in K), and $P = (0 : 0 : 1)$ is singular on $C : y^2z = x^3$. □

In order to define the group structure on an elliptic curve, we need a statement about the intersection of a line and a curve.

22.8. **Proposition.** *Let $L : ax + by + cz = 0$ be a line and $C : F(x, y, z) = 0$ a projective curve of degree d such that $L \not\subset C$ (i.e., F is not divisible by $ax + by + cz$). Then for every sufficiently large field $K' \supset K$ (we can take K' to be an algebraic closure of K), the intersection $L(K') \cap C(K')$ has exactly d points, counting multiplicity.*

Proof. Here is a sketch of a proof. Not all of a, b, c are zero, so let us assume without loss of generality that $c \neq 0$. Then we can solve the equation of L for z:

$$z = -\frac{a}{c}\, x - \frac{b}{c}\, y$$

and plug this into the equqation of C:

$$F_1(x, y) = F\Big(x, y, -\frac{a}{c}\, x - \frac{b}{c}\, y\Big) = 0\,,$$

where F_1 is a homogeneous polynomial of degree d in two variables. (Note that $F_1 \neq 0$ since there are points on L on which F does not vanish.) Therefore F_1 splits into linear factors over some finite field extension K' of K:

$$F_1(x, y) = \gamma \prod_{i=1}^{k} (\alpha_i x - \beta_i y)^{e_i}$$

with $\alpha_i, \beta_i, \gamma \in K'$, $\alpha_i \beta_j \neq \alpha_j \beta_i$ for $i \neq j$, $e_i \geq 1$, and $\sum_{i=1}^{k} e_i = d$. The points in $L \cap C$ are then

$$P_i = (c\beta_i : c\alpha_i : -a\beta_i - b\alpha_i) \in L(K') \cap C(K')\,,$$

and the statement is true if we count P_i with multiplicity e_i. □

22.9. **Remark.** If, in the situation above, $d - 1$ of the intersection points are defined over K (i.e., are in $\mathbb{P}^2(K)$), then so is the last, dth, one. The reason for this is that (assuming for simplicity that y does not divide F_1) $F_1(x, 1) \in K[x]$ is a polynomial of degree d such that $d-1$ of its roots are in K, and therefore the last root must be in K as well (the sum of the roots is minus a quotient of coefficients of $F_1(x, 1)$ and therefore in K).

Using this remark, we can define the group law (as we did earlier).

22.10. **Theorem.** *Let E be an elliptic curve over K. Then the set $E(K)$ of K-rational points on E has the structure of an abelian group with zero element $O = (0 : 1 : 0)$ and addition*

$$P + Q = (P * Q) * O \qquad \textit{(and negation} \quad -P = P * O\textit{)}$$

where $P * Q$ denotes the third point of intersection of the line through P and Q (the tangent line to E at P when $P = Q$) with E.

We will not give a full proof here, but we note that all axioms other than associativity are easily verified. To show associativity, it is enough to show that

$$(P + Q) * R = P * (Q + R)$$

for all triples of points $P, Q, R \in E(K)$. If the points are in sufficiently general position, one can make use of the following fact.

22.11. **Theorem.** *Let L_1, L_2, L_3 and L'_1, L'_2, L'_3 be six lines in sufficiently general position. Let P_{ij} be the point of intersection of L_i and L'_j. Then every cubic curve that passes through P_{ij} for all i, j except $i = j = 3$ also passes through P_{33}.*

Proof. Here is a sketch. To pass through a given point P imposes a linear condition on the coefficients of a general homogeneous cubic polynomial in three variables. The "sufficiently general position" ensures that the eight conditions we obtain are linearly independent (this is essentially the definition of "sufficiently general position"). Since a general homogeneous polynomial of degree 3 has ten coefficients, the space of cubics passing through the eight points has dimension two. Now (writing L_1 etc. for the linear form defining L_1 etc.) $L_1L_2L_3$ and $L'_1L'_2L'_3$ are two such cubics, and they are linearly independent. So any F giving a curve passing through the eight points is a linear combination of these two cubics, and therefore F vanishes also on P_{33}. □

Now we take $L_1 = L(P, Q)$, $L_2 = L(O, Q*R)$, $L_3 = L(P+Q, R)$ and $L'_1 = L(Q, R)$, $L'_2 = L(O, P * Q)$, $L'_3 = L(P, Q + R)$, where $L(X, Y)$ denotes the line through X and Y. Then by the theorem above, we find that E passes through P_{33}, which is therefore the third point of intersection of E with L_3 and with L'_3. This means that $(P + Q) * R = P * (Q + R)$, as was to be shown.

Note that this proof does not work as described if some of the nine points coincide, and one either has to consider a whole lot of special cases separately, or use some sort of "continuity argument" to get rid of them.

22.12. **Definition.** Let E be an elliptic curve given by a "short Weierstrass Equation"

$$y^2 = x^3 + A\,x + B\,.$$

Then the number $\Delta(E) = -16(4A^3 + 27B^2)$ is called the *discriminant* of E.

22.13. **Remark.** The number $-4A^3 - 27B^2$ is the discriminant of the polynomial $f(x) = x^3 + Ax + B$, which is defined to be

$$\operatorname{disc}(f) = (\alpha - \beta)^2(\beta - \gamma^2)(\gamma - \alpha)^2\,,$$

where α, β, γ are the three roots of f. Therefore the discriminant vanishes if and only if f has multiple roots.

The strange factor 16 makes things work in characteristic 2: one can define $\Delta(E)$ for "long Weierstrass Equations"

$$E : y^2 + a_1\,xy + a_3\,y = x^3 + a_2\,x^2 + a_4\,x + a_6$$

by formally transforming it into a short one (using a substitution $x \mapsto x + \alpha$, $y \mapsto y + \beta x + \gamma$) and then taking Δ of the result. With the factor 16, $\Delta(E)$ is a polynomial in the coefficients $a_1, \ldots, a_6$, with integral coefficients, and vanishes if and only if there is a singularity, also when the characteristic is 2 or 3.

As an example, consider $y^2 - y = x^3 - x$. What would be the discriminant? We complete the square on the left and obtain $(y - 1/2)^2 = x^3 - x + 1/4$. Replacing $y - 1/2$ by a new y, we have a short Weierstrass Equation, and its discriminant is $-16(4 \cdot (-1)^3 + 27 \cdot (1/4)^2) = 37$. Since this is not divisible by 2, the original equation defines an elliptic curve over any field of characteristic 2 (in fact, any field of characteristic different from 37).

22.14. **The rational torsion group.** In an abelian group, the subset of elements of finite order forms a subgroup. Let E be an elliptic curve, then a point P on E is called a *torsion point*, if $nP = O$ for some $n \geq 1$. The smallest such n is then the *order* of P (as usual for group elements). When G is an abelian group, we write

$$G[n] = \{g \in G : ng = 0\}$$

for the *n-torsion subgroup* of elements killed by n and $G_{\text{tors}} = \bigcup_{n\geq 1} G[n]$ for the torsion subgroup of G.

What can the torsion subgroup of $E(\mathbb{Q})$ look like? It is a fact that as groups,

$$E(\mathbb{C}) \cong (\mathbb{R}/\mathbb{Z})^2 \quad \text{and} \quad E(\mathbb{R}) \cong \mathbb{R}/\mathbb{Z} \text{ or } \mathbb{Z}/2\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$$

(the first when $x^3 + Ax + B$ has only one real root, the second when it has three). This implies that any finite group of rational torsion points must be of the form $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2n\mathbb{Z}$.

The following result tells us that $E(\mathbb{Q})_{\text{tors}}$ is finite and therefore must be of one of these types.

22.15. **Theorem (Nagell-Lutz).** *Let E be given by $y^2 = x^3 + A\,x + B$, where A and B are integers. Then for every torsion point $P = (\xi, \eta) \in E(\mathbb{Q}) \setminus \{O\}$:*

(1) *ξ and η are integers.*

(2) *$\eta = 0$ or η^2 divides $4A^3 + 27B^2$.*

In particular, $E(\mathbb{Q})_{tors}$ is finite: there are only finitely many possible y-coordinates, and for each y-coordinate, there are at most three corresponding x-coordinates.

Furthermore, this result provides us with an algorithm to find all the torsion points: first, we find all integral points (x, y) on E such that y^2 divides $4A^3 + 27B^2$. Then for each such point P, we compute its multiples nP, $n = 2, 3, \ldots$, until we either obtain O (then the point is a torsion point, and we will also have found its order), or we obtain a point that does not belong to the set of points we have determined in the first step (then P must have infinite order).

Proof. (Sketch) We begin with the easy part, and this is that the first statement implies the second. Assume P is a torsion point such that $2P \neq O$ (then $P = (\xi, \eta)$ with $\eta \neq 0$). We have to show that η^2 divides $4A^3 + 27B^2$. Now observe that $2P$ is also a torsion point (and $\neq O$), so by the first statement, its x-coordinate ξ' is integral. It is easy to compute

$$\xi' = \frac{\xi^4 - 2A\xi^2 - 8B\xi + A^2}{4(\xi^3 + A\xi + B)}\,.$$

Consider the following trivially verified equality:

$$(3x^2+4A)(x^4-2Ax^2-8Bx+A^2)-(3x^3-5Ax-27B)(x^3+Ax+B) = 4A^3+27B^2$$

We plug in ξ for x and divide by $\eta^2 = \xi^3 + A\xi + B \neq 0$ to get

$$\frac{4A^3 + 27B^2}{\eta^2} = 4(3\xi^2 + 4A)\xi' - (3\xi^3 - 5A\xi - 27B) \in \mathbb{Z}\,.$$

It is quite a bit harder to prove the first part. First observe the following.

A point $P = (\xi, \eta) \in E(\mathbb{Q}) \setminus \{O\}$ has the form

$$P = \left(\frac{r}{t^2}, \frac{s}{t^3}\right) \quad \text{where } r, s, t \in \mathbb{Z} \text{ with } r \perp t\,, s \perp t\,.$$

Proof. We look at one prime p at a time and have to show that

$$v_p(\xi) < 0 \iff v_p(\eta) < 0 \iff v_p(\xi) = -2\nu\,, v_p(\eta) = -3\nu \text{ with } \nu > 0\,.$$

Now if $v_p(\xi) < 0$, then $2v_p(\eta) = v_p(\xi^3 + A\xi + B) = 3v_p(\xi) < 0$, so $v_p(\eta) < 0$ and $2v_p(\eta) = 3v_p(\xi)$, which implies the last statement. If $v_p(\xi) \geq 0$, then $v_p(\eta) \geq 0$, so $v_p(\eta) < 0$ implies $v_p(\xi) < 0$. This completes the proof. □

The basic idea is now this. Suppose that $P = (\xi, \eta)$ is not integral. Then $\xi = r/t^2$, $\eta = s/t^3$, and $t > 1$, so there is some prime p dividing t. Let $\nu = v_p(t)$. Then I claim that $pP = (r'/t'^2, s'/t'^3)$ with $v_p(t') = \nu + 1$. This implies that the denominators (of the x-coordinates, say) of the points $P, pP, p^2P, \ldots$ are all distinct, hence the points are all distinct, hence P cannot have finite order (because then it would only have finitely many distinct multiples).

The idea for proving the claim is that in the p-adic metric, a large ν in the above means that ξ and η are large, and hence P is close to O. Then the claim says that multiplying P by p moves it closer to O.

For a point $P = (\xi, \eta) \in E(\mathbb{Q})$ such that $2P \neq O$, we define $w(P) = \xi/\eta$. If $v_p(\xi) = -2\nu < 0$ as above, then $v_p(w(P)) = \nu$. It is now possible (and not very hard) to show that if $P_1, P_2 \in E(\mathbb{Q})$ are points with p dividing the denominators of their x-coordinates and such that $v_p(w(P_j)) \geq \nu$ $(j = 1, 2)$, then

$$w(P_1) + w(P_2) \equiv w(P_1 + P_2) \bmod p^{5\nu}\,.$$

By induction, this implies, for $w(P) = \nu > 0$,

$$w(mP) \equiv mw(P) \bmod p^{5\nu}\,,$$

and from this we get

$$v_p(w(p^nP)) = v_p(w(P)) + n = \nu + n\,,$$

as claimed above. □

22.16. **Examples.** Consider the curve $E : y^2 = x^3 + 1$. We find $4A^3 + 27B^2 = 27$. Hence all the torsion points $P \neq O$ have

$$y(P) \in \{0, \pm 1, \pm 3\}\,.$$

For $y = 0$, we find $x = -1$; this gives a point $(-1, 0)$ of order 2. For $y = \pm 1$, we find $x = 0$, and $(0, \pm 1)$ is a triple intersection point of E with the line $y = \pm 1$. This implies that these points have order 3. The sum $(-1, 0) + (0, 1)$ must then have order 6, and indeed, for $y = 3$, we find $x = 2$ and the points $(2, \pm 3)$ of order 6.

Note that in general, we have to test the integral points we find whether they really are torsion points. For example, on $y^2 = x^3 + 3x$, there is the integral point $P = (1, 2)$ the square of whose y-coordinate divides $4A^3 + 27B^2 = 108$. We find $2P = (1/4, -7/8)$, therefore P cannot be a torsion point (and this proves that there are infinitely many rational points on that curve). Let us see what the torsion subgroup is in this example. The other possibilities for the y-coordinate are $0, \pm 1, \pm 3, \pm 6$. We find the point $(0, 0)$ of order 2, but no other integral points.

Having proved that for each individual elliptic curve, the rational torsion subgroup is finite, one can ask whether the order of this group is uniformly bounded (the alternative would be that the group can be arbitrarily large). The answer to this question is yes.

22.17. **Theorem (Mazur).** If E is an elliptic curve over $\mathbb{Q}$, then the torsion subgroup $E(\mathbb{Q})_{\text{tors}}$ is one of the following 15 groups.

$$\begin{aligned} \mathbb{Z}/n\mathbb{Z} \quad & \text{for } n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12 \quad \text{or} \\ \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2n\mathbb{Z} \quad & \text{for } n = 1, 2, 3, 4. \end{aligned}$$

All of these groups occur (even in a one-parameter family of elliptic curves).

The proof of this theorem is far beyond what we can do in this course.

22.18. **Reduction mod p.** In general, if we have a Weierstrass equation with integral coefficients, it makes sense to ask whether we obtain an elliptic curve over the finite field $\mathbb{F}_p$ if we reduce the coefficients mod p. Since the discriminant is a polynomial with integral coefficients in the coefficients of the Weierstrass equation, we obtain its value for the mod p reduced equation by reducing the discriminant mod p. This implies that we get an elliptic curve E_p over $\mathbb{F}_p$ if and only if the discriminant is not divisible by p. In this case, we say that p is a *prime of good reduction* for our elliptic curve.

In this case, we have the group $E(\mathbb{Q})$ on the one side and the finite group $E_p(\mathbb{F}_p)$ on the other side. Is there any relation between the two? Ideally, we would like to take a point in $E(\mathbb{Q})$ and reduce its coordinates mod p to get a point in $E_p(\mathbb{F}_p)$. There may be a problem, however, when p divides the denominator of the coordinates. To see how we can avoid this problem, let us consider reduction mod p of points in the projective plane.

22.19. **Lemma.** *There is a canonical map $\rho_p : \mathbb{P}^2(\mathbb{Q}) \to \mathbb{P}^2(\mathbb{F}_p)$, which is defined as follows. Let $P = (\xi : \eta : \zeta) \in \mathbb{P}^2(\mathbb{Q})$. Then we can scale the coordinates to get $P = (\xi' : \eta' : \zeta')$ with coprime integers ξ', η', ζ'. Then $\rho_p(P) = (\bar{\xi}' : \bar{\eta}' : \bar{\zeta}')$.*

It is easy to check that the map is well-defined: the only ambiguity in the triple (ξ', η', ζ') is a simultaneous sign change, which leads to the same point $\rho_p(P)$. Also, since ξ', η', ζ' are coprime, at least one of $\bar{\xi}$, $\bar{\eta}'$ and $\bar{\zeta}'$ will be non-zero.

We see that the use of projective coordinates eliminates problems with denominators; this is another advantage of the projective plane over the affine plane.

We can now restrict ρ_p to the points on our elliptic curve.

22.20. **Proposition.** *Let E be an elliptic curve over $\mathbb{Q}$, given by an equation with integral coefficients, and let p be a prime of good reduction for E (i.e., such that $p \nmid \Delta_E$). Then*

$$\rho_p : E(\mathbb{Q}) \longrightarrow E_p(\mathbb{F}_p)$$

is a group homomorphism.

It is clear that this is well-defined as a map. In order to show that it is a group homomorphism, one needs to check that the points of intersection of E with a line are mapped to the points of intersection of E_p with a line, with multiplicities behaving as they should. This is not hard.

Let $P = (r/t^2, s/t^3)$ be in the kernel of ρ_p. To see what this means, we first have to write P as a point with projective coordinates that are coprime integers. We get $P = (rt : s : t^3)$. This reduces to $\bar{O} = (\bar{0} : \bar{1} : \bar{0})$ if and only if $\bar{t} = \bar{0}$, i.e., if and only if p divides t. We see that $\ker \rho_p$ is the subgroup of $E(\mathbb{Q})$ consisting of O and the points whose coordinates have denominators divisible by p.

22.21. **Corollary.** *If $E : y^2 = x^3 + Ax + B$ with $A, B \in \mathbb{Z}$, and $p \nmid \Delta_E$, then the homomorphism ρ_p is injective on $E(\mathbb{Q})_{\text{tors}}$.*

Proof. By the Nagell-Lutz Theorem 22.15, every point $P \in E(\mathbb{Q})_{\text{tors}} \setminus \{O\}$ has integral coordinates, hence $P \notin \ker \rho_p$. So $\ker \rho_p \cap E(\mathbb{Q})_{\text{tors}} = \{O\}$. □

This can be used to bound the size of the torsion subgroup: ρ_p identifies it with a subgroup of $E_p(\mathbb{F}_p)$; therefore $\#E(\mathbb{Q})_{\text{tors}}$ must divide $\#E_p(\mathbb{F}_p)$.

22.22. **Example.** Consider $E : y^2 - y = x^3 - x$. We saw earlier that $\Delta_E = 37$. Counting points, we find $\#E_2(\mathbb{F}_2) = 5$ and $\#E_3(\mathbb{F}_3) = 7$. So the order of $E(\mathbb{Q})_{\text{tors}}$ must divide both 5 and 7, therefore $E(\mathbb{Q})_{\text{tors}} = \{O\}$.

Integral Points. We know that there can be infinitely many rational points on an elliptic curve. What about integral points?

22.23. **Example.** Consider $E : y^2 = x^3 + 17$. We find the following integral points:

$$(-2, \pm 3)\,, \quad (-1, \pm 4)\,, \quad (2, \pm 5)\,, \quad (4, \pm 9)$$
$$(8, \pm 23)\,, \quad (43, \pm 282)\,, \quad (52, \pm 375)\,, \quad (5234, \pm 378661)$$

The example shows that there can be quite a number of integral points, and it is clear that this number is not bounded: take any elliptic curve with infinitely many rational points, then by scaling the variables suitably, we can make as many of them integral as we like. So the question remains whether any given elliptic curve can have infinitely many integral points.

22.24. **Theorem (Siegel).** *Let E be an elliptic curve over $\mathbb{Q}$, given by an equation with integral coefficients. Then E has only finitely many integral points.*

This is not easy to prove. Siegel's proof is not *effective:* it does not provide a bound for the coordinates of the integral points. It also does not lead to an algorithm that determines all integral points. Note that the difficulty is not to *find* the points, but to know when we have found *all* of them. What Siegel proves is roughly that when there is a very large integral point, then there can be no integral point that is still very much larger. This leads to a contradiction when we assume that there are infinitely many integral points. But it does not tell us that the list we have produced is complete (unless we really find a very large point).

Later, Baker found an effective bound (using his bounds on linear forms in logarithms of algebraic numbers); however this bound is really huge (something like doubly exponential in the coefficients). So this is a big theoretical improvement, but it does not lead to a practical algorithm.

However, if we know generators of the group $E(\mathbb{Q})$, then there is a method that, given a bound, produces a new bound that usually is smaller. Iterating this, one arrives at a bound that is manageable for most reasonable example cases. So in practice, we usually *can* find the set of integral points. For example, it can be shown that the points we listed above in Example 22.23 are all the integral points on $y^2 = x^3 + 17$.

The Group of Rational Points. The last result I would like to discuss in this section is the fundamental structure theorem for the group of rational points on an elliptic curve.

22.25. **Theorem (Mordell-Weil).** *If E is an elliptic curve over $\mathbb{Q}$, then the group $E(\mathbb{Q})$ is a finitely generated abelian group.*

The result as stated was proved by Mordell in a paper from 1922. Weil later generalized it to higher-dimensional "abelian varieties" (one-dimensional abelian varieties are the same as elliptic curves) and arbitrary number fields (i.e., finite field extensions of $\mathbb{Q}$) instead of $\mathbb{Q}$.

By the general result on the structure of finitely generated abelian groups, this implies that as abelian groups,

$$E(\mathbb{Q}) \cong E(\mathbb{Q})_{\text{tors}} \oplus \mathbb{Z}^r$$

for some $r \geq 0$, which is called the *(Mordell-Weil) rank* of $E(\mathbb{Q})$. It is an open problem (but widely believed to be true) whether r can be arbitrarily large. The latest record example I know of had $r \geq 24$, but nobody so far was able to produce a sequence of elliptic curves whose ranks tend to infinity.

The next question is if and how we can actually determine r for a given elliptic curve. This is also an open problem: there is no method for which one could prove that it determines the rank in all cases. There are, however, methods that work in practice for more or less all examples (with reasonably-sized coefficients). If one could prove another standard conjecture ("the Shafarevich-Tate group of E is finite"), then (at least in principle) these methods would find the rank eventually.

Now let me briefly indicate how the theorem is proved. The first step is to prove the following result (which clearly must hold if $E(\mathbb{Q})$ is finitely generated).

22.26. **Theorem (Weak Mordell-Weil Theorem).** *The group $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite.*

Proof. (Sketch) I will give the idea of the proof in a special case, namely that all the points of order 2 on E are rational. Then E has an equation of the form

$$E : y^2 = (x-a)(x-b)(x-c)$$

with integers a, b, c, and the three points of order 2 are $(a, 0)$, $(b, 0)$ and $(c, 0)$. Note that a, b, c are pairwise distinct.

The idea of the proof is to embed $E(\mathbb{Q})/2E(\mathbb{Q})$ into a group which can be shown to be finite. To do this, we define a map

$$\begin{aligned}
\varphi : \quad E &\longrightarrow \mathbb{Q}^\times/(\mathbb{Q}^\times)^2 \times \mathbb{Q}^\times/(\mathbb{Q}^\times)^2 \times \mathbb{Q}^\times/(\mathbb{Q}^\times)^2 \\
O &\longmapsto (1,1,1) \\
(\xi,\eta) &\longmapsto (\xi-a, \xi-b, \xi-c) \quad \text{if } \xi \neq a,b,c \\
(a,0) &\longmapsto ((a-b)(a-c), a-b, a-c) \\
(b,0) &\longmapsto (b-a, (b-a)(b-c), b-c) \\
(c,0) &\longmapsto (c-a, c-b, (c-a)(c-b))
\end{aligned}$$

(the values given are representatives of the classes mod $(\mathbb{Q}^\times)^2$). Now I make a number of claims.

(1) If $\varphi(P) = (\alpha, \beta, \gamma)$, then $\alpha\beta\gamma$ is a square (i.e., $\alpha\beta\gamma = 1$ in $\mathbb{Q}^\times/(\mathbb{Q}^\times)^2$).
(2) φ is a group homomorphism.
(3) $\ker\varphi = 2E(\mathbb{Q})$.
(4) $\varphi(E(\mathbb{Q})) \subset H \times H \times H$, where H is the subgroup of $\mathbb{Q}^\times/(\mathbb{Q}^\times)^2$ that is generated by the classes of -1, 2, and the prime numbers p dividing $(b-a)(c-b)(a-c)$.

Note that H is finite and hence so is $H \times H \times H$. Statements (2) and (3) then show that φ induces an injection of $E(\mathbb{Q})/2E(\mathbb{Q})$ into the finite group $H \times H \times H$, and the statement of the theorem follows (in the special case considered).

The proofs of these claims are not very hard, but too lengthy to do them here.

In the general case (with non-rational 2-torsion points), one still uses the same idea, but one has to work with the number fields obtained by adjoining roots of $x^3 + Ax + B$ to $\mathbb{Q}$. This requires some knowledge of Algebraic Number Theory (which studies such fields). This proof generalizes to the case that E is an elliptic curve over any number field. □

22.27. **Remark.**

(1) Since by the first claim in the proof above, the product of the three components of $\varphi(P)$ is always 1 (in $\mathbb{Q}^\times/(\mathbb{Q}^\times)^2$), the first two components determine the last one uniquely. This implies that we even obtain an injection of $E(\mathbb{Q})/2E(\mathbb{Q})$ into $H \times H$.

(2) If r is the rank of $E(\mathbb{Q})$, then $E(\mathbb{Q}) = T \oplus \mathbb{Z}^r$ with a finite group $T = E(\mathbb{Q})_{\text{tors}}$. Then

$$E(\mathbb{Q})/2E(\mathbb{Q}) = T/2T \oplus (\mathbb{Z}/2\mathbb{Z})^r \cong T[2] \oplus (\mathbb{Z}/2\mathbb{Z})^r\,,$$

where $T[2] = E(\mathbb{Q})[2] = \{P \in E(\mathbb{Q}) : 2P = O\}$ is the 2-torsion subgroup of T (and of $E(\mathbb{Q})$). Hence

$$\#(E(\mathbb{Q})/2E(\mathbb{Q})) = \#T[2] \cdot 2^r = 2^{r+2}$$

(the last equality is only valid in the special case considered in the proof). This then implies that

$$r \le 2 + 2\#\{p : p \text{ odd}, p \mid (a-b)(b-c)(c-a)\} = 2 + 2s$$

(note that $\#H = 2^{2+s}$ and $2^{2+r} \le \#H^2$).
Exercise. If T is a finite abelian group and p is a prime number, then $\#(T/pT) = \#T[p]$ (which implies that there is a (non-canonical) isomorphism between T/pT and $T[p]$).

(3) By using "local information" (signs or p-adic considerations for the "bad" primes p), it is possible to restrict the image of φ further. For example, if $a < b < c$, then the first component of $\varphi(P)$ is always positive, which means that the image of φ is contained in a subgroup of index 2 of $H \times H$. This improves the bound given above to

$$r \le 1 + 2s\,.$$

22.28. **Example.** Consider

$$E : y^2 = x^3 - x = (x+1)x(x-1)\,.$$

The only prime dividing one of the differences of the roots of the right hand side is 2. Therefore, H in the proof above is generated by (the classes of) -1 and 2, and s above is zero. From Remark 22.27 above, we see that the image of φ is contained in

$$\{(\alpha,\beta,\gamma) \in \langle -1, 2\rangle^3 : \alpha\beta\gamma = 1, \alpha > 0\}\,,$$

a group of order 8, and the rank of $E(\mathbb{Q})$ is bounded by 1. Now we consider to what extent 2 can show up in α, β and γ. One can check (this is part of the argument in the proof of claim (4) in the proof of the weak Mordell-Weil Theorem 22.26) that points that have even denominators in their x-coordinate will map to elements

that do not involve 2 (the 2-adic valuations are even). Otherwise, $x(P)$ is either even (even numerator) or odd (odd numerator). In the first case, $x(P)\pm 1$ are both odd, hence $v_2(\alpha) = v_2(\gamma) = 0$, which implies that $v_2(\beta)$ is even (by claim (1)), hence β does not involve 2. In the second case, $v_2(\beta) = v_2(x(P)) = 0$. We see that β can only be ± 1, hence

$$\varphi(E(\mathbb{Q})) \subset \{(\alpha, \beta, \alpha\beta) : \alpha \in \langle 2\rangle, \beta \in \langle -1\rangle\}\,.$$

This group has order 4. But we know

$$\varphi((-1,0)) = (2,-1,-2)\,, \quad \varphi((0,0)) = (1,-1,-1)\,, \quad \varphi((1,0)) = (2,1,2)\,.$$

This means that we have equality above, and the rank is zero, so $E(\mathbb{Q})$ is finite. (Since $\#\varphi(E(\mathbb{Q})) \geq \#E(\mathbb{Q})[2]$, we know that we must have equality even without computing images of points!)

In order to determine $E(\mathbb{Q})$ completely, we note that all possibly existing additional points are torsion points and therefore must be integral, with the square of the y-coordinate dividing $4A^3 + 27B^2 = 4$, see the Nagell-Lutz Theorem 22.15. But neither $x^3 - x - 1$ nor $x^3 - x - 4$ have integral roots. Hence we can conclude that

$$E(\mathbb{Q}) = \{O, (-1,0), (0,0), (1,0)\}\,.$$

To conclude the proof of the Mordell-Weil Theorem, we need to establish a way of measuring the "size" of a point in $E(\mathbb{Q})$. Then we can apply the following result.

22.29. **Lemma.** *Suppose that G is an abelian group and that $h : G \to \mathbb{R}_+$ is a map satisfying the following assumptions.*

(1) *$G/2G$ is finite.*

(2) *For each $Q \in G$, there is some $C_Q \geq 0$ such that*

$$h(P+Q) \leq 2\,h(P) + C_Q \quad \text{for all } P \in G.$$

(3) *There is a constant $C \geq 0$ such that*

$$h(2P) \geq 4\,h(P) - C \quad \text{for all } P \in G.$$

(4) *For every $B > 0$, the set $\{P \in G : h(P) \leq B\}$ is finite.*

Then G is finitely generated.

Proof. By the first assumption, we can pick a finite set $S \subset G$ of coset representatives of $2G$. Let $D = \max\{C_{-Q} : Q \in S\}$. We show that G is in fact generated by S, together with all elements P such that $h(P) \leq (C+D)/2$. Call this set T. By the last assumption, this generating set T is finite, which proves the lemma.

Assume the claim is false. Then there is some $P \in G$ such that P is not in the subgroup generated by T. By the last assumption, we can asume that $h(P)$ is minimal among all such elements. Trivially, we must have $h(P) > (C+D)/2$. Now, there is some $Q \in S$ such that $P - Q = 2P' \in 2G$. Then we have

$$h(P') \leq (h(2P') + C)/4 \leq (2h(P) + C_{-Q} + C)/4 \leq (2h(P) + C + D)/4 < h(P)\,.$$

By our choice of P, P' is in the subgroup generated by T, but then $P = Q + 2P'$ must be in there as well, a contradiction. □

22.30. **The height.** We have proved so far that $G = E(\mathbb{Q})$ satisfies the first assumption (at least when all the 2-torsion points are rational). So it remains to define a suitable map h in order to finish the proof of the Mordell-Weil Theorem.

Let $P = (r/t^2, s/t^3) \in E(\mathbb{Q})$, where $r \perp t$, $s \perp t$. Then we define

$$h(P) = \log \max\{|r|, t^2\} \quad \text{and} \quad h(O) = 0\,.$$

(This is called the *logarithmic naive height* on $E(\mathbb{Q})$.)

We have to show that this map h satisfies the asumptions in Lemma 22.29. The last one is the easiest one to see: When $h(P)$ is bounded, then numerator and denominator of $x(P)$ are bounded (by $e^{h(P)}$), so there are only finitely many possible x-coordinates. But for each x-coordinate, there are at most two points in $E(\mathbb{Q})$, hence any set of points of bounded height is finite.

22.31. **Lemma.** There is a constant C such that for all $P, Q \in E(\mathbb{Q})$,

$$|h(P+Q) + h(P-Q) - 2\,h(P) - 2\,h(Q)| \le C\,.$$

We skip the proof. Since $h(P-Q) \ge 0$, this provides us with the middle two assumptions in Lemma 22.29. Therefore, all the assumptions of Lemma 22.29 are satisfied. Hence $E(\mathbb{Q})$ is finitely generated.

22.32. **Discussion.** How can we try to determine $E(\mathbb{Q})$? The first step is to bound the order of $E(\mathbb{Q})/2E(\mathbb{Q})$ as tightly as possible, by embedding it into a finite group like $H \times H$ in the proof of Thm. 22.26 and making use of all the restrictions we can derive from considerations at bad primes, compare Example 22.28. Then we hope to find sufficiently many points in $E(\mathbb{Q})$ to show that our bound is actually tight. If we are successful, then we know the rank of $E(\mathbb{Q})$. What is more, we also know generators of a subgroup of finite odd index (by taking a set of points whose images under φ generate the image of φ). We can then use refinements of the height estimates used in the lemma above to find actual generators of the free part $E(\mathbb{Q})$. The torsion subgroup can be dealt with using the Nagell-Lutz result 22.15.

If we are not successful in establishing that the bound on $E(\mathbb{Q})/2E(\mathbb{Q})$ is tight, then this can be for two reasons. It is possible that the bound is tight, but that we just have not found some of the preimages (usually this happens because these points are simply too large to be found by search). But it is also possible that the bound is *not* tight. To make progress in this situation, one can write down equations whose solutions parametrize the points in the preimage under φ of a given element of the group bounding the image of φ. Then one can search for solutions of these equations; the advantage here is that these solutions are (usually) smaller than the points in $E(\mathbb{Q})$ they give rise to, hence they are found more easily. This often helps resolve the first case (when the points are too large). It does not help in the second case, when there are no solutions to be found for some of the equations. In this case, one can try to iterate the procedure that produced these equations (also known as "first descents") to produce so-called "second descents". Sometimes, this enables one to deduce that a first descent does not have a solution (because it does not produce any second descents). In other cases, one can use the second descents (the solutions of which are again smaller) to produce solutions to a first descent.

An alternative approach is to play a similar game with $E(\mathbb{Q})/3E(\mathbb{Q})$ (or $E(\mathbb{Q})/5E(\mathbb{Q})$ or ...). Working out methods for doing this or second (and third, ...) descents is the subject of current active research.

23. Primes in arithmetic progressions

Euclid showed that there are infinitely many primes. His argument can be refined to give more information, for instance:

23.1. **Theorem.** There are infinitely many primes congruent to 3 mod 4.

Proof. Suppose not, and so suppose that $p_1, \ldots, p_k$ is the complete list of primes congruent to 3 mod 4. Form their product; if that is 1 mod 4, add 2. Otherwise add 4. The resulting number n is congruent to 3 mod 4, and is not divisible by any p_i. Hence all prime factors of n must be 1 mod 4, which implies $n \equiv 1 \bmod 4$. This contradiction completes the proof. □

To show that there are infinitely many primes $p \equiv 1 \bmod 4$, one can use that an odd prime p is 1 mod 4 if and only if -1 is a quadratic residue mod p. So, taking

$$P = 4(p_1 p_2 \cdots p_k)^2 + 1\,,$$

$P \geq 5$ is not divisible by any of the p_j, and all its prime divisors are $\equiv 1 \bmod 4$. By the same kind of argument as above, this implies that there are infinitely many primes $p \equiv 1 \bmod 4$.

Exercise: Prove in a similar spirit that there are infinitely many primes $p \equiv 1 \bmod 2^n$, for every $n \geq 1$.

It is true, however, that all residue classes contain infinitely many primes, except when there is an obvious reason why not.

23.2. **Theorem (Dirichlet).** *Suppose N and a are coprime integers. Then there are infinitely many primes congruent to a modulo N.*

Dirichlet proved this in the 1830s or so, and the rest of this section discusses his proof. Dirichlet's approach grows out of an exotic proof that there are infinitely many primes given by Euler a century earlier. Consider the function

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}$$

defined on the set of complex numbers s with $\operatorname{Re} s > 1$. This is the *Riemann zeta function*, as it has been known since "Riemann's memoir" from 1860, in which Riemann outlined a strategy for proving the *prime number theorem*, which predicts how frequently prime numbers occur among positive integers, using complex analytic properties of $\zeta(s)$. That is a story for another section.

Note that for $\operatorname{Re} s > 1$, the series defining $\zeta(s)$ converges absolutely. Now, observe the formal identity

$$\prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1} = \prod_{p \text{ prime}} \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \ldots\right) = \sum_{n=1}^{\infty} \frac{1}{n^s}\,,$$

which expresses the fact that each positive integer n can be written as a product of primes in exactly one way. For each $\operatorname{Re} s > 1$, all the sums and products converge absolutely, so one may change the order of the terms, and therefore both sides

converge to the same value. (In fact, both sides define the same analytic function on $\operatorname{Re} s > 1$.)

Recall that $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. Equivalently, $\lim_{s\to 1^+} \zeta(s) = +\infty$, since for real values of s, all the terms in $\sum_{n=1}^{\infty} \frac{1}{n^s}$ are positive. Hence,

$$\lim_{s\to 1^+} \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1} = +\infty\,,$$

which implies there must be infinitely many primes. This was essentially Euler's proof. Of course, one can also get some quantitative information out of it. One sees that $\prod_{p \text{ prime}} \left(1 - \frac{1}{p}\right)^{-1}$ must diverge, which implies by a calculation that $\sum_{p \text{ prime}} \frac{1}{p}$ must diverge (see below). This would not be the case if, for instance, the primes were distributed as thinly as the squares. In fact, since $\sum_{n=1}^{\infty} \frac{1}{n^{1+\epsilon}}$ converges for any fixed real number $\epsilon > 0$, we must have

$$\#\{\text{prime } p < N\} > N^{1-\epsilon}\,,$$

for infinitely many different integers N. We will learn more about the distribution of primes in the next section.

The Strategy. Let us now discuss the strategy for the proof of the general statement in Thm. 23.2. The proof mentioned above that there are infintely many can be phrased like this. Consider

$$\begin{aligned}\log \zeta(s) = \log \prod_p \frac{1}{\left(1 - \frac{1}{p^s}\right)} &= \sum_p \log \frac{1}{\left(1 - \frac{1}{p^s}\right)} \\ &= \sum_p \left(\frac{1}{p^s} + \frac{1}{2p^{2s}} + \dots\right) = \sum_p \frac{1}{p^s} + f(s)\end{aligned}$$

where $f(s)$ remains bounded as $s \to 1^+$:

$$f(s) = \sum_p \sum_{k=2}^{\infty} \frac{1}{kp^{ks}} \le \frac{1}{2} \sum_p \frac{1}{p^2} \frac{1}{1 - 1/p} \le \sum_p \frac{1}{p^2} \le \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

Now, as $s \to 1^+$, $\zeta(s) \to \infty$, and therefore $\sum_p p^{-s} \to \infty$ as well, showing that there are infintely many primes (and even that $\sum_p 1/p$ diverges).

Now the idea is to show in a similar way that

$$\sum_{p \equiv a \bmod N} p^{-s} \to \infty \quad \text{as } s \to 1^+.$$

However, there is no simple way to set this up directly, using some modification of $\zeta(s)$, so we have to use a slight detour. We basically only can hope to prove something about functions like $\zeta(s)$ if they involve all the primes. We also want these functions to have a similar product structure (so that we can take logarithms easily). This leads to the following notion.

23.3. **Definition.** A *Dirichlet character* mod N is a function $\chi : \mathbb{Z} \to \mathbb{C}$ with the following properties.

(1) $\chi(n)$ only depends on the residue class of n mod N.
(2) $\chi(n) \neq 0$ if and only if $n \perp N$.
(3) $\chi(mn) = \chi(m)\chi(n)$ for all $m, n \in \mathbb{Z}$.

The *L-series* associated to χ is defined for $s > 1$ by

$$L(\chi, s) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = \prod_p \Big(1 - \frac{\chi(p)}{p^s}\Big)^{-1}.$$

(The second equality is seen in the same way as for $\zeta(s)$, using the multiplicativity of χ. A product representation of this type is called an *Euler product*.)

23.4. **Example.** For $N = 4$, a Dirichlet character χ mod 4 is determined by its values at 1 and 3 (the values at even numbers must be zero). The multiplicativity forces $\chi(1) = 1$. Since $3^2 \equiv 1 \bmod 4$, we must have $\chi(3)^2 = \chi(1) = 1$, so $\chi(3) = \pm 1$. One checks easily that both choices give a Dirichlet character mod 4. Call them χ_0 and χ_1, respectively. Then

$$L(\chi_0, s) = 1 + \frac{1}{3^s} + \frac{1}{5^s} + \cdots = \prod_{p \nmid 4} \Big(1 - \frac{1}{p^s}\Big)^{-1} = \Big(1 - \frac{1}{2^s}\Big)\,\zeta(s)$$

and

$$L(\chi_1, s) = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \dots .$$

Note that (by the alternating series criterion) $L(\chi_1, s)$ converges for real $s > 0$ (uniformly for $s \geq \delta > 0$), and $L(\chi_1, 1) = \arctan 1 = \pi/4 \neq 0$.

It is a fact (which we will prove soon) that

$$\frac{1}{\phi(N)} \sum_{\chi} \overline{\chi(a)}\chi(n) = \begin{cases} 1 & \text{if } a \equiv n \bmod N \\ 0 & \text{else} \end{cases}$$

(where χ runs through all the Dirichlet characters mod N.) This means that we can take suitable linear combinations of $\log L(\chi, s)$ in order to single out the residue class we are interested in. We have

$$\log L(\chi, s) = \sum_p \frac{\chi(p)}{p^s} + f_\chi(s)$$

as before, where again f_χ remains bounded as $s \to 1^+$. So

$$\frac{1}{\phi(N)} \sum_{\chi} \overline{\chi(a)} \log L(\chi, s) = \sum_{p \equiv a \bmod N} \frac{1}{p^s} + f_a(s)$$

with f_a bounded as $s \to 1^+$. In order to finish the proof, we have to show that the left hand side tends to infinity as $s \to 1^+$.

23.5. **Example.** For $N = 4$ again, we find

$$\frac{1}{2}\big(\log L(\chi_0, s) + \log L(\chi_1, s)\big) = \sum_{p \equiv 1 \bmod 4} p^{-s} + \text{bounded}$$

$$\frac{1}{2}\big(\log L(\chi_0, s) - \log L(\chi_1, s)\big) = \sum_{p \equiv 3 \bmod 4} p^{-s} + \text{bounded}$$

We have seen that $L(\chi_1, s)$ is a continuous function for real $s > 0$ that does not vanish for $s \geq 1$. This implies that $\log L(\chi_1, s)$ stays bounded as $s \to 1^+$. Since $L(\chi_0, s)$ is essentially $\zeta(s)$, $\log L(\chi_0, s)$ tends to $+\infty$ as $s \to 1^+$. This proves again that there are infinitely many primes $p \equiv 1 \bmod 4$ and infinitely many primes $p \equiv 3 \bmod 4$.

Like χ_0 in the example above, there is always one special Dirichlet character mod N, the so-called *trivial* Dirichlet character χ_0, which takes the value 1 on all numbers coprime with N (and the value 0 otherwise). It is easy to see that its L-series is, up to a simple factor, $\zeta(s)$, and so

$$\log L(\chi_0, s) \to \infty \quad \text{as } s \to 1^+.$$

It remains to prove that the other summands cannot lead to any cancellation. This is done by showing that

(1) the series defining $L(\chi, s)$ converges for $s > 0$, hence the function is defined (and continuous) there, and
(2) $L(\chi, 1) \neq 0$.

These statements imply that $\log L(\chi, s) \to \log L(\chi, 1)$ stays bounded as $s \to 1^+$ for $\chi \neq \chi_0$, concluding the proof.

Now we have to fill in the various details.

Characters. Let us first deal with the Dirichlet characters. There is a general notion of characters, as follows.

23.6. **Definition.** Let G be a group. A *character* of G is a group homomorphism $\chi : G \to \mathbb{C}^\times$. Let $\hat{G}$ be the set of all characters of G; then $\hat{G}$ is an abelian group (the *character group* of G) under point-wise multiplication:

$$(\chi\psi)(g) = \chi(g)\psi(g)\,.$$

23.7. **Remark.** There is a bijection between Dirichlet characters mod N and the character group of $(\mathbb{Z}/N\mathbb{Z})^\times$, as follows. If χ is a Dirichlet character mod N, then $\psi(\bar{a}) := \chi(a)$ defines a character on $(\mathbb{Z}/N\mathbb{Z})^\times$ (well-defined since $\chi(a)$ only depends on $\bar{a}$, and for $\bar{a} \in (\mathbb{Z}/N\mathbb{Z})^\times$, $\chi(a) \neq 0$). If ψ is a character of $(\mathbb{Z}/N\mathbb{Z})^\times$, then $\chi(a) = \psi(\bar{a})$ for $a \perp N$, $\chi(a) = 0$ otherwise, defines a Dirichlet character mod N.

Now for finite abelian groups (like $(\mathbb{Z}/N\mathbb{Z})^\times$), the character group behaves particularly nicely.

23.8. **Proposition.** *If G is a finite abelian group, then its character group $\hat{G}$ is isomorphic to G.*

Proof. Assume first that G is cyclic of order n, and pick a generator g. Then for any character χ, we must have $\chi(g^k) = \chi(g)^k$; therefore, χ is completely determined by the value $\chi(g)$. Now there is only one relation for $g \in G$, namely $g^n = 1$. Therefore, $\chi(g)^n = 1$ as well, and every choice of $\chi(g)$ satisfying this will define a character. This shows that

$$\mu_n \longrightarrow \hat{G}, \quad \zeta \longmapsto (g^k \mapsto \zeta^k)$$

sets up an isomorphism of $\hat{G}$ with the group μ_n of nth roots of unity, which is a cyclic group of order n, hence isomorphic to G. (Note that the isomorphism $G \cong \hat{G}$ we construct depends on choices of generators of G and of μ_n: it is not canonical.)

Now we prove that $\widehat{G \times H} \cong \hat{G} \times \hat{H}$. The isomorphism is given as follows. Let $\chi \in \hat{G}$ and $\psi \in \hat{H}$. then $\phi(g, h) = \chi(g)\psi(h)$ defines a character of $G \times H$. Conversely, if ϕ is a character of $G \times H$, then $\phi|_G$ and $\phi|_H$ are characters of G

and H, respectively, and one easily checks that the two maps are inverses of each other. (Note that this isomorphism *is* canonical; it does not depend on choices.)

Finally, we use that every finite abelian group is a direct product of cyclic groups:

$$G = G_1 \times \cdots \times G_k \cong \hat{G}_1 \times \cdots \times \hat{G}_k \cong (G_1 \times \cdots \times G_k)\hat{} = \hat{G}$$

□

For example, this implies that there are exactly $\#(\mathbb{Z}/N\mathbb{Z})^\times = \phi(N)$ distinct Dirichlet characters mod N. Next we want to state and prove the "orthogonality relations" we need for the proof of Dirichlet's Theorem 23.2.

23.9. **Lemma.** *Let G be a finite abelian group and $1 \neq g \in G$. Then there is a character $\chi \in \hat{G}$ such that $\chi(g) \neq 1$.*

Proof. First assume G is cyclic. Then we can take any character χ that generates $\hat{G}$. (Compare the preceding proof.)

In the general case, write $G = G_1 \times \cdots \times G_k$ as a product of cyclic groups. Then $g = (g_1, \ldots, g_k)$ with at least one $g_j \neq 1$. Take a character χ_j of G_j such that $\chi_j(g_j) \neq 1$, and define $\chi(h_1, \ldots, h_k) = \chi_j(h_j)$. □

23.10. **Proposition.** *Let G be a finite abelian group. Then*

(1) *For all $\chi \in \hat{G}$,*

$$\sum_{g \in G} \chi(g) = \begin{cases} \#G & \text{if } \chi = 1_{\hat{G}} \\ 0 & \text{else} \end{cases}$$

and therefore for all $\chi_1, \chi_2 \in \hat{G}$

$$\frac{1}{\#G} \sum_{g \in G} \overline{\chi_1(g)} \chi_2(g) = \begin{cases} 1 & \text{if } \chi_1 = \chi_2 \\ 0 & \text{else} \end{cases}$$

(2) *For all $g \in G$,*

$$\sum_{\chi \in \hat{G}} \chi(g) = \begin{cases} \#G & \text{if } g = 1_G \\ 0 & \text{else} \end{cases}$$

and therefore for all $g_1, g_2 \in G$

$$\frac{1}{\#G} \sum_{\chi \in \hat{G}} \overline{\chi(g_1)} \chi(g_2) = \begin{cases} 1 & \text{if } g_1 = g_2 \\ 0 & \text{else} \end{cases}$$

Proof. If $\chi = 1_{\hat{G}}$, then $\chi(g) = 1$ for all $g \in G$, and the first claim is trivial. Otherwise, there is an $h \in G$ such that $\chi(h) \neq 1$. Then

$$(1 - \chi(h)) \sum_g \chi(g) = \sum_g \chi(g) - \sum_g \chi(hg) = \sum_g \chi(g) - \sum_g \chi(g) = 0\,,$$

and so $\sum_g \chi(g) = 0$ as well (since $1 - \chi(h) \neq 0$). To get the assertion on χ_1 and χ_2, apply the result to $\chi = \overline{\chi}_1 \chi_2 = \chi_1^{-1} \chi_2$.

If $g = 1_G$, then $\chi(g) = 1$ for all $\chi \in \hat{G}$, and the second claim is trivial (using $\#\hat{G} = \#G$). Otherwise, by Lemma 23.9, there is a $\psi \in \hat{G}$ such that $\psi(g) \neq 1$. Then

$$(1 - \psi(g)) \sum_\chi \chi(g) = \sum_\chi \chi(g) - \sum_\chi (\psi\chi)(g) = \sum_\chi \chi(g) - \sum_\chi \chi(g) = 0\,,$$

and so $\sum_\chi \chi(g) = 0$. To get the assertion on g_1 and g_2, apply this to $g = g_1^{-1} g_2$ and note that $\chi(g_1^{-1}) = \chi(g_1)^{-1} = \overline{\chi(g_1)}$. □

23.11. **Corollary.** *Applying the preceding orthogonality relation to Dirichlet characters, we get for $a \perp N$ that*

$$\frac{1}{\phi(N)} \sum_\chi \overline{\chi(a)}\chi(n) = \begin{cases} 1 & \text{if } a \equiv n \bmod N \\ 0 & \text{else} \end{cases}$$

where the sum is over the Dirichlet characters mod N.

This is one of the ingredients needed in the proof of Dirichlet's Theorem.

The other essential part of the proof is to deal with the behavior of the functions $\log L(\chi, s)$ as $s \to 1^+$. Let us first consider the trivial character χ_0 (it is the Dirichlet character corresponding to $1_{\widehat{\mathbb{Z}/N\mathbb{Z}^\times}}$).

23.12. **Lemma.** *Let χ_0 be the trivial Dirichlet character mod N. Then for $s > 1$,*

$$L(\chi_0, s) = \prod_{p|N} \Big(1 - \frac{1}{p^s}\Big) \zeta(s)$$

and therefore

$$\log L(\chi_0, s) - \log \zeta(s) = \sum_{p|N} \log(1 - p^{-s})$$

is bounded as $s \to 1^+$.

Proof. Compare the formulas

$$\zeta(s) = \prod_p (1 - p^{-s})^{-1}$$

and

$$L(\chi_0, s) = \prod_p (1 - \chi_0(p) p^{-s})^{-1} = \prod_{p \nmid N} (1 - p^{-s})^{-1}\,.$$

□

Before we can proceed, we need a basic result on the convergence properties of "Dirichlet series" $\sum_{n=1}^\infty a_n n^{-s}$.

23.13. **Proposition.**

(1) *Let (a_n), (b_n) be two sequences of complex numbers, and let $A_n = \sum_{k=1}^n a_k$. Then*

$$\sum_{n=1}^N a_n b_n = \sum_{n=1}^{N-1} A_n (b_n - b_{n+1}) + A_N b_N\,.$$

(2) *Keeping the notations, assume that* (A_n) *is bounded and that* (b_n) *is a decreasing sequence of real numbers such that* $b_n \to 0$. *Then* $\sum_{n=1}^{\infty} a_n b_n$ *converges, and*

$$\sum_{n=1}^{\infty} a_n b_n = \sum_{n=1}^{\infty} A_n (b_n - b_{n+1}) \,.$$

(3) *Let* $f(s) = \sum_{n=1}^{\infty} c_n n^{-s}$. *If the series converges for* $s = s_0$, *then it converges uniformly for all* $s \geq s_0 + \delta$, *for any* $\delta > 0$, *and defines an analytic function on* $s > s_0$.

Proof. (1) This is an easy exercise.
(2) Let $|A_n| \leq A$ for all n. Then for $M < N$, we have by (1)

$$\sum_{n=M+1}^{N} a_n b_n = \sum_{n=M}^{N-1} A_n (b_n - b_{n+1}) + A_N b_N - A_M b_M \,.$$

Hence

$$\Big| \sum_{n=M+1}^{N} a_n b_n \Big| \leq A \Big(\sum_{n=M}^{N-1} (b_n - b_{n+1}) + b_M + b_N \Big) = A(b_M - b_N + b_M + b_N) = 2A b_M \to 0$$

as $M \to \infty$, hence the sequence of partial sums is Cauchy. Since $A_N b_N \to 0$, we get the stated formula by taking limits in (1).
(3) Set $a_n = c_n n^{-s_0}$ and $b_n = n^{s_0 - s}$. Then the assumptions in (2) are satisfied for $s \geq s_0 + \delta$ (since A_n converges to $f(s_0)$). From the proof of (2), we get the uniform bound

$$\Big| \sum_{n=M+1}^{N} c_n n^{-s} \Big| \leq 2A M^{-\delta} \,,$$

which shows uniform convergence. Now a uniformly convergent series of analytic functions converges to an analytic function (compare Introductory Complex Analysis), and since we can take $\delta > 0$ as small as we like, we get a function that is analytic on $s > s_0$. □

The following result tells us more precisely what the behavior of $\zeta(s)$ is near $s = 1$.

23.14. **Lemma.** *There is an analytic function* $f(s)$ *for* $s > 0$ *such that for* $s > 1$,

$$\zeta(s) = \frac{1}{s-1} + f(s) \,.$$

We define $\zeta(s)$ *for* $0 < s < 1$ *by this formula.*

Proof. First note that

$$F(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^s} = 1 - \frac{1}{2^s} + \frac{1}{3^s} - \frac{1}{4^s} + - \dots$$

defines an analytic function for $s > 0$: by part (2) in Prop. 23.13 (with $a_n = (-1)^{n-1}$ and $b_n = n^{-s}$), the series converges for $s > 0$, and by part (3) it gives an analytic function there. Now for $s > 1$,

$$\Big(1 - \frac{2}{2^s}\Big)\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \dots - \frac{2}{2^s} - \frac{2}{4^s} - \dots = F(s) \,.$$

Therefore,

$$\zeta(s) = \frac{1}{1-2^{1-s}}F(s) = \frac{1}{s-1}\cdot\frac{s-1}{1-2^{1-s}}F(s)$$
$$= \frac{1}{s-1}\Big(\frac{F(1)}{\log 2} + (s-1)f(s)\Big) = \frac{1}{s-1} + f(s)$$

with an analytic function $f(s)$; note that $F(1) = 1-1/2+1/3-+\cdots = \log 2$. □

23.15. **Corollary.** *We have*

$$\log\zeta(s) = \log\frac{1}{s-1} + \text{bounded} \quad \textit{as } s \to 1^+.$$

Now we look at the L-series $L(\chi, s)$.

23.16. **Proposition.** *Let χ be a nontrivial Dirichlet character mod N. Then the series*

$$L(\chi, s) = \sum_{n=1}^{\infty}\frac{\chi(n)}{n^s}$$

converges for $s > 0$ and defines therefore an analytic function on this domain.

Proof. We apply Prop. 23.13 again. Set $a_n = \chi(n)$ and $b_n = n^{-s}$. Then $|A_n| \le \phi(N)$, since $\sum_{n=kN+1}^{(k+1)N}\chi(n) = 0$ for all k. Hence the assumptions in part (2) are satisfied, and the argument proceeds as for $F(s)$ in the proof of the lemma above. □

It remains to show that $L(\chi, 1) \neq 0$, for then $\log L(\chi, s)$ will stay bounded as $s \to 1^+$. This is the hardest part in the proof of Dirichlet's Theorem 23.2. We will state an auxiliary result first.

23.17. **Theorem (Landau).** *Suppose that $a_n \ge 0$ for all n and that the series $f(s) = \sum_{n=1}^{\infty} a_n n^{-s}$ converges for some, but not for all $s \in \mathbb{R}$. Let*

$$s_0 = \inf\{s \in \mathbb{R} : \textit{the series converges}\}.$$

Then $f(s)$ cannot be continued to the left of s_0 as an analytic function.

We will not prove this result here; it makes use of the fact that $f(s)$ is even analytic for $\operatorname{Re} s > s_0$. (Here is a sketch: Under the assumption that $f(s)$ is analytic on a neighborhood of s_0, the power series expansion of $f(s)$ around $s_0 + 1$ has radius of convergence > 1; convergence of this at $s_0 - \delta$ then implies convergence of the series definig $f(s)$ there: a contradiction.)

As an example, consider $\zeta(s)$: here $s_0 = 1$, and $\zeta(s)$ has a pole at $s = 1$ (it can be extended to a meromorphic function on $\mathbb{C}$ with just this one pole, though). On the other hand, for

$$\log\zeta(s) = \sum_p\sum_{k=1}^{\infty}\frac{1}{kp^{ks}},$$

we have again $s_0 = 1$, but this time, there is no meromorphic continuation across s_0. As a last example, consider $L(\chi, s)$ for a non-trivial Dirichlet character χ. Then $s_0 = 0$, but one can show that $L(\chi, s)$ extends to an entire function. So the conclusion of the theorem is false in this case, showing that the assumption $(a_n \ge 0)$ is essential.

To finish the proof of Dirichlet's Theorem 23.2, we need one more lemma.

23.18. **Lemma.** *Let $F(s) = \prod_\chi L(\chi, s)$, where the product is over all Dirichlet characters mod N. Then for $s > 1$,*

$$F(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s},$$

where $a_n \geq 0$ for all n and $a_n \geq 1$ for $n = m^{\phi(N)}$ with some $m \geq 1$, $m \perp N$.

Proof. We first need a formula about characters: Let G be a finite abelian group, and lat $g \in G$ be an element of order f. Then

$$\prod_{\chi \in \hat{G}} (1 - \chi(g)X) = (1 - X^f)^{\#G/f}$$

as polynomials in X. To see this, we observe that $\hat{G} \ni \chi \longmapsto \chi(g) \in \mu_f$ is a surjective group homomorphism. If we let $\zeta = \exp(2\pi i/f)$ be a generator of μ_f, then it follows that

$$\prod_{\chi \in \hat{G}} (1 - \chi(g)X) = \prod_{k=0}^{f-1} (1 - \zeta^k X)^{\#G/f} = (1 - X^f)^{\#G/f}.$$

Now we consider the Euler product of $F(s)$ (everything converges absolutely for $s > 1$, justifying the rearrangements):

$$F(s) = \prod_\chi L(\chi, s) = \prod_\chi \prod_p (1 - \chi(p)p^{-s})^{-1} = \prod_p \Big(\prod_\chi (1 - \chi(p)p^{-s}) \Big)^{-1}$$

and by the result above:

$$= \prod_{p \nmid N} (1 - p^{-f_p s})^{\phi(N)/f_p} = \prod_{p \nmid N} (1 + p^{-f_p s} + p^{-2f_p s} + \dots)^{\phi(N)/f_p}$$

In each factor of the last product, all terms have nonnegative coefficients, and all terms of the form $c_k p^{-k\phi(N)s}$ have coefficient ≥ 1. The claim follows by expanding the product. □

Now we can finish the proof.

23.19. **Theorem.** *If χ is a non-trivial Dirichlet character mod N, then we have $L(\chi, 1) \neq 0$.*

Proof. Assume that $L(\chi, 1) = 0$ for some (non-trivial) χ. Then the simple pole of $L(\chi_0, s)$ at $s = 1$ is canceled by the zero of $L(\chi, s)$ at $s = 1$; therefore the function $F(s) = \prod_\chi L(\chi, s)$ is analytic for $s > 0$. But by the preceding lemma, we have that $F(s) = \sum_{n=1}^\infty a_n n^{-s}$ with $a_n \geq 0$. Furthermore, $a_{m^{\phi(N)}} \geq 1$ for all $m \perp N$, and so the series does not converge for $s = 1/\phi(N)$:

$$\sum_{n=1}^{\infty} a_n n^{-1/\phi(N)} \geq \sum_{m \geq 1, m \perp N} (m^{\phi(N)})^{-1/\phi(N)} = \sum_{m \geq 1, m \perp N} \frac{1}{m} = \infty$$

So by Landau's Theorem 23.17, $F(s)$ has no analytic continuation across s_0, where $1/\phi(N) \leq s_0 \leq 1$, a contradiction. So our assumption that $L(\chi, 1) = 0$ for some χ must be false. □

24. The Prime Number Theorem

The distribution of the prime numbers among the integers is a mystery that was (and is) studied by many mathematicians for a long time. Let $\pi(x)$ (for $x \in \mathbb{R}$) denote the number of primes $p \le x$. After extensive computations of tables of primes, Gauss conjectured around 1800 (and others, like Legendre, did the same at about the same time) that the "density" of primes near x should be $1/\log x$. More precisely, he conjectured that for large x,

$$\pi(x) \sim \mathrm{li}(x) = \mathrm{li}(2) + \int_2^x \frac{dt}{\log t}$$

(here, li(2) is defined by the Cauchy principal value of the integral:

$$\mathrm{li}(2) = \lim_{\varepsilon \to 0^+} \int_0^{1-\varepsilon} \frac{dt}{\log t} + \int_{1+\varepsilon}^2 \frac{dt}{\log t} \approx 1.045)$$

This notation means that

$$\lim_{x\to\infty} \frac{\pi(x)}{\mathrm{li}(x)} = 1\,.$$

Legendre's version was

$$\pi(x) \sim \frac{x}{\log x}\,;$$

integrating by parts, it is easy to see that the two versions are equivalent.

It took nearly 100 years before this statement, known as the *Prime Number Theorem* was finally proved. The most important input came again from Riemann's memoir of 1860, where he introduced the study of the (now so-called) *Riemann zeta function* $\zeta(s)$ as a function of a complex argument s.

About ten years earlier, Chebyshev could at least prove that the order of magnitude was correct: there are constants $0 < c < C$ such that for $x \ge 2$ (say),

$$c\,\frac{x}{\log x} \le \pi(x) \le C\,\frac{x}{\log x}\,.$$

The argument is rather elementary. First we define two more functions that are easier to deal with. In the following, p always denotes a prime number. Let

$$\theta(x) = \sum_{p\le x} \log p$$

and (the first sum runs over pairs (p,k))

$$\psi(x) = \sum_{p^k\le x} \log p = \sum_{p\le x} \Big\lfloor \frac{\log x}{\log p} \Big\rfloor \log p = \sum_{n\le x} \Lambda(n) = \sum_{k=1}^{\infty} \theta(x^{1/k})\,.$$

Here, the *von Mangoldt function* Λ is defined as

$$\Lambda(p^k) = \log p\,, \quad \Lambda(n) = 0 \text{ otherwise}\,.$$

Now we have that

$$\theta(x) \le \psi(x) = \sum_{p\le x} \Big\lfloor \frac{\log x}{\log p} \Big\rfloor \log p \le \pi(x)\,\log x\,.$$

Also, for $\varepsilon > 0$ small,

$$\theta(x) \ge \sum_{x^{1-\varepsilon}<p\le x} \log p \ge (\pi(x) - \pi(x^{1-\varepsilon}))(1-\varepsilon)\log x \ge (1-\varepsilon)\log x(\pi(x) - x^{1-\varepsilon})$$

Now let us consider the prime factorization of the binomial coefficient $\binom{2n}{n}$.

24.1. **Lemma.** We have

$$\theta(2n) - \theta(n) \le \log \binom{2n}{n} \le \psi(2n)\,.$$

Proof. Let $n < p \le 2n$ be a prime. Then p divides the numerator of $\binom{2n}{n} = (2n)!/n!^2$ once, but does not divide the denominator. Hence the binomial coefficient is divisible by $\prod_{n<p\le 2n} p$. This implies the first inequality.

For the second inequality, observe that

$$v_p\Big(\binom{2n}{n}\Big) = \sum_{k=1}^{\infty}\Big(\Big\lfloor \frac{2n}{p^k}\Big\rfloor - 2\Big\lfloor \frac{n}{p^k}\Big\rfloor\Big) \le \Big\lfloor \frac{\log(2n)}{\log p}\Big\rfloor\,.$$

□

Since $\binom{2n}{n} \sim 4^n/\sqrt{\pi n}$ by Stirling's formula, we get

$$\psi(2n) \ge 2n\log 2 - O(\log n)$$

and

$$\theta(2^k) = \sum_{j=1}^{k}(\theta(2^j) - \theta(2^{j-1})) \le \sum_{j=1}^{k} 2^k \log 2 \le 2^k \cdot 2\log 2\,.$$

These imply

$$\psi(x) \ge \log 2 \cdot x - O(\log x)$$

and

$$\theta(x) \le \theta(2^{\lceil \log x/\log 2\rceil}) \le 2\log 2 \cdot 2^{\lceil \log x/\log 2\rceil} \le 4\log 2 \cdot x\,.$$

We already see that

$$\pi(x) \ge \frac{\psi(x)}{\log x} \ge \log 2\,\frac{x}{\log x} - O(1)\,.$$

If we set $\varepsilon = 2\log\log x/\log x$ above, we obtain

$$\pi(x) \le \frac{\theta(x)}{\log x}\Big(1 + O\Big(\frac{\log\log x}{\log x}\Big)\Big) \le 4\log 2\,\frac{x}{\log x}\Big(1 + O\Big(\frac{\log\log x}{\log x}\Big)\Big)\,.$$

However, in order to prove the Prime Number Theorem completely, more is needed. The ground-breaking ideas were formulated by Riemann in his memoir *Über die Anzahl der Primzahlen unter einer gegebenen Größe* (On the number of primes below a given quantity), relating the asymptotics of $\pi(x)$ with analytic properties of $\zeta(s)$ as a meromorphic function. However, it still took more than 30 years, until Hadamard and De la Vallée Poussin independently were able to carry through the program sketched by Riemann.

24.2. **Lemma.** We have

$$\lim_{x\to\infty}\Big(\frac{\psi(x)}{x} - \frac{\pi(x)}{x/\log x}\Big) = 0\,.$$

Proof. We have already seen that

$$(1-\varepsilon)\log x(\pi(x) - x^{1-\varepsilon}) \le \theta(x) \le \psi(x) \le \pi(x)\log x\,.$$

We divide by x, set $\varepsilon = 2\log\log x/\log x$ and let x tend to infinity; the result follows. □

To prove the prime number theorem, it therefore suffices to show that

$$\psi(x) \sim x\,, \quad \text{i.e.,} \quad \lim_{x\to\infty} \frac{\psi(x)}{x} = 1\,.$$

The following lemma shows why it is advantageous to work with $\psi(x)$ instead of $\pi(x)$.

24.3. **Lemma.** For $\operatorname{Re} s > 1$, we have

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}\,.$$

Proof. We compute the logarithmic derivative using the Euler product for $\zeta(s)$. (Note that for $\operatorname{Re} s > 1$, everything converges absolutely and locally uniformly, hence all manipulations with the series below are justified.)

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_p \frac{d}{ds} \log(1-p^{-s}) = \sum_p \log p\, \frac{p^{-s}}{1-p^{-s}} = \sum_p \sum_{k=1}^{\infty} \frac{\log p}{p^{ks}} = \sum_n \frac{\Lambda(n)}{n^s}$$

□

Now it is easy to see that $\zeta(s)$ extends to a meromorphic function on $\operatorname{Re} s > 0$ with only a simple pole at $s = 1$. (In fact, $\zeta(s)$ extends to a meromorphic function of all of $\mathbb{C}$, with still only this one simple pole.) Hence $-\zeta'(s)/\zeta(s)$ likewise is a meromorphic function there, with simple poles at the pole of $\zeta(s)$ (residue 1) and the zeros of $\zeta(s)$ (residue $-$the order of the zero).

If one would want to work directly with $\pi(x)$, the corresponding function is $\sum_p p^{-s}$, which does not have nice properties. One could try to use $\log\zeta(s) = \sum_p \sum_{k=1}^{\infty} p^{-ks}/k$ (the difference between $\pi(x)$ and the corresponding function for $\log\zeta(s)$ is negligible), but this function has logarithmic singularities at $s = 1$ and at the zeros of $\zeta(s)$, which complicates things considerably.

24.4. **Proposition (Perron's Formula and an Application).**

(1) Let $c > 0$, $y > 0$. Then

$$\lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} \frac{y^s}{s}\, ds = \begin{cases} 0 & \text{if } 0 < y < 1, \\ \frac{1}{2} & \text{if } y = 1, \\ 1 & \text{if } y > 1. \end{cases}$$

(2) Let $f(s) = \sum_{n=1}^{\infty} a_n\, n^{-s}$, and suppose that the series converges absolutely for $s = c > 0$. Then for $x > 0$,

$$\lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} f(s)\, \frac{x^s}{s}\, ds = \begin{cases} \sum\limits_{n\le x} a_n & \text{if } x \notin \mathbb{Z}, \\ \sum\limits_{n<x} a_n + \frac{1}{2} a_x & \text{if } x \in \mathbb{Z}. \end{cases}$$

Proof. I will not give the details of the proof. The idea for part (1) is to move the line segment $[c - iT, c + iT]$ far off to the right (if $0 < y < 1$) or to the left (if $y > 1$). In the first case, the integral around the rectangle one gets is zero by the residue theorem, and the integrals over the three sides different from $[c-iT, c+iT]$ can be estimated to tend to zero (as the right hand vertical edge moves to infinity and $T \to \infty$). In the second case, the argument is similar, but here, the integral

around the rectangle picks up the residue of y^s/s at $s = 0$, which is 1. The case $y = 1$ is done by direct calculation.

To prove (2), one applies (1); one has to justify the swapping of the sum with the integral and limit, but this can be done using the estimates one obtains in the proof of (1) and the assumption that the series converges absolutely for $s = c$.

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} f(s)\,\frac{x^s}{s}\,ds &= \lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} \sum_{n=1}^{\infty} a_n\,\frac{(x/n)^s}{s}\,ds \\ &= \sum_{n=1}^{\infty} a_n \lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} \frac{(x/n)^s}{s}\,ds \\ &= \sum_{n=1}^{\infty} a_n \begin{cases} 1 & \text{if } n < x, \\ \frac{1}{2} & \text{if } n = x, \\ 0 & \text{if } n > x. \end{cases}\end{aligned}$$

□

24.5. **Corollary.** Let $c > 1$. Then for $x > 0$,

$$\lim_{T\to\infty} \frac{1}{2\pi i} \int_{c-iT}^{c+iT} \Big(-\frac{\zeta'(s)}{\zeta(s)}\,\frac{x^s}{s}\Big)\,ds = \begin{cases} \psi(x) & \text{if } x \notin \mathbb{Z}, \\ \psi(x) - \frac{1}{2}\Lambda(x) & \text{if } x \in \mathbb{Z}. \end{cases}$$

Proof. Clear. □

So in order to prove the Prime Number Theorem $\psi(x) \sim x$, we have to evaluate/estimate the integral. The idea is to move the line over which we integrate a little bit to the left (at least some part of it with bounded imaginary part). Then the integral will pick up the residue of $-\zeta'(s)x^s/(\zeta(s)s)$ at $s = 1$, which is x. But this will only work if we do not pick up any other poles. This in turn means that $\zeta(1 + it) \neq 0$ for all (real) $t \neq 0$ (otherwise, there will be poles on $\operatorname{Re} s = 1$, and we cannot move our integration contour across this line).

24.6. **Lemma.** $\zeta(s)$ does not vanish on $\operatorname{Re} s = 1$.

Proof. First we show that for $\sigma > 1$ and $t \in \mathbb{R}$, we have

$$|\zeta(\sigma)^3\zeta(\sigma + it)^4\zeta(\sigma + 2it)| \geq 1\,.$$

To see this, we take the logarithm of the left hand side:

$$\begin{aligned}\log|\zeta(\sigma)^3\zeta(\sigma + it)^4\zeta(\sigma + 2it)| &= \operatorname{Re}\big(3\log\zeta(\sigma) + 4\log\zeta(\sigma + it) + \log\zeta(\sigma + 2it)\big) \\ &= \sum_p \sum_{k=1}^{\infty} \frac{1}{kp^{k\sigma}} \operatorname{Re}(3 + 4e^{-kti\log p} + e^{-2kti\log p}) \\ &= \sum_p \sum_{k=1}^{\infty} \frac{1}{kp^{k\sigma}} (3 + 4\cos(kt\log p) + \cos(2kt\log p)) \\ &\geq 0\end{aligned}$$

since $3 + 4\cos\alpha + \cos 2\alpha = 2(1 + \cos\alpha)^2 \geq 0$.

Now assume that $\zeta(1+it)=0$ for some $t\neq 0$. Then $\zeta(\sigma+it)/(\sigma-1)$ remains bounded as $\sigma\to 1^+$; also $\zeta(\sigma)(\sigma-1)$ is bounded as $\sigma\to 1^+$. Hence

$$1\le |\zeta(\sigma)^3\zeta(\sigma+it)^4\zeta(\sigma+2it)| = \Big|\big(\zeta(\sigma)(\sigma-1)\big)^3\Big(\frac{\zeta(\sigma+it)}{\sigma-1}\Big)^4\zeta(\sigma+2it)\Big|(\sigma-1)\to 0$$

as $\sigma\to 1^+$, a contradiction. □

To finish the proof of the Prime Number Theorem, one has to obtain a suitable "zero-free region" for $\zeta(s)$ to the left of $\operatorname{Re} s=1$ and a bound on $\zeta'(s)/\zeta(s)$ there. This is a bit technical (though not really hard), and we will omit the details here.

What one gets fairly easily is that $\zeta(\sigma+it)$ does not vanish for $\sigma>1-C/\max\{1,\log|t|\}$ for some constant C; this then translates into an error term in the Prime Number Theorem of the form

$$\psi(x)=x+O(xe^{-c\sqrt{\log x}})\quad\text{or}\quad \pi(x)=\operatorname{li}(x)+O(xe^{-c\sqrt{\log x}})\,.$$

(There are some improvements on this, replacing the square root by a higher power of $\log x$, like $(\log x)^{3/5-\varepsilon}$.)

24.7. **Final Remarks.** Here are some more facts about the Riemann zeta function $\zeta(s)$.

(1) $\zeta(s)$ can be extended to a meromorphic function on $\mathbb{C}$; its only pole is the simple pole at $s=1$ with residue 1.

(2) $\zeta(s)$ has simple zeros at negative even integers ("trivial zeros"); it also has infinitely many zeros in the "critical strip" $0<\operatorname{Re} s<1$.

(3) $\zeta(s)$ satisfies a functional equation: define

$$\xi(s)=s(1-s)\pi^{-s/2}\Gamma\Big(\frac{s}{2}\Big)\zeta(s)\,;$$

then $\xi(s)$ is an entire function satisfying $\xi(1-s)=\xi(s)$. Its zeros coincide with the nontrivial zeros of $\zeta(s)$.
($\Gamma(s)$ is the Gamma function; it can be defined for $\operatorname{Re} s>0$ by the integral

$$\Gamma(s)=\int_0^\infty e^{-t}t^{s-1}\,dt\,;$$

it satisfies the functional equation $s\Gamma(s)=\Gamma(s+1)$, which can be used to define it as a meromorphic function on all of $\mathbb{C}$. It has no zeros, and has simple poles at the nonpositive integers. Also $\Gamma(1)=1$ and therefore (by induction), $\Gamma(n+1)=n!$.)

(4) Riemann (in his 1860 memoir) stated it was "likely" that all the nontrivial zeros are actually on the "critical line" $\operatorname{Re} s=1/2$. This is the famous *Riemann Hypothesis*.

Using this (and suitable estimates for $\zeta'(s)/\zeta(s)$ when $\operatorname{Re} s\ll 0$), one can continue moving the line of integration to the left and finally obtain the *Explicit Formula*

$$\psi(x)=x-\sum_\rho\frac{x^\rho}{\rho}-\log 2\pi-\log(1-x^{-2})\,.$$

Here, ρ runs through the zeros of $\zeta(s)$ in the critical strip, with multiplicities, and the sum has to be taken in ascending order of $|\operatorname{Im}\rho|$. (Note that $\zeta'(0)/\zeta(0)=\log 2\pi$.) There is a similar formula for $\pi(x)$.

From this explicit formula or the idea of the proof of the Prime Number Theorem, one obtains the following.

24.8. **Proposition.** Let $1/2 \le \beta < 1$.

(1) If $\zeta(s)$ does not vanish for $\operatorname{Re} s > \beta$, then
$$\psi(x) = x + O(x^\beta (\log x)^2) \quad \text{and} \quad \pi(x) = \operatorname{li}(x) + O(x^\beta (\log x)^2)\,.$$
(2) If
$$\psi(x) = x + O(x^\beta) \quad \text{or} \quad \pi(x) = \operatorname{li}(x) + O(x^\beta)\,,$$
then $\zeta(s)$ does not vanish for $\operatorname{Re} s > \beta$.

As a corollary, we see that the Riemann Hypothesis is equivalent to the purely number theoretic statement
$$\psi(x) = x + O(x^{1/2+\varepsilon}) \quad \text{for all } \varepsilon > 0.$$
In some sense, this expresses that the distribution of the prime numbers is as even as possibly allowed by the fact that $\zeta(s)$ does have infinitely many zeros in the critical strip.

Another view on this is the following. Let $\mu(n)$ be the *Möbius function*:
$$\mu(n) = \begin{cases} (-1)^r & \text{if } n = p_1 \dots p_r \text{ is squarefree } (r \ge 0), \\ 0 & \text{otherwise.} \end{cases}$$
Then it is easy to see that $\sum_{n=1}^\infty \mu(n) n^{-s} = 1/\zeta(s)$. Now, applying the above approach to $1/\zeta(s)$ instead of $-\zeta'(s)/\zeta(s)$, we obtain that the Riemann Hypothesis is equivalent to
$$\sum_{n \le x} \mu(n) = O(x^{1/2+\varepsilon}) \quad \text{for all } \varepsilon > 0.$$
This can be interpreted as saying that the sequence of the nonzero values of $\mu(n)$ behaves statistically like an unbiased random walk.

References

[Bru] J. Brüdern: *Einführung in die analytische Zahlentheorie.* Springer Verlag, 1995.

[Cas] J.W.S. Cassels: *Lectures on elliptic curves.* LMS Student Texts **24**, Cambridge University Press, 1991.

[IrRo] K. Ireland, M. Rosen: *A classical introduction to modern number theory.* Springer Graduate Texts in Mathematics 84, Springer Verlag, 1982.

[Nat] M.B. Nathanson: *Elementary methods in number theory.* Springer Graduate Texts in Mathematics 195, Springer Verlag, 2000.

[Sil] Joseph H. Silverman: *The Arithmetic of elliptic curves.* Springer Verlag, 1986.

[Sti] Douglas R. Stinson: *Cryptography. Theory and practice.* 2nd edition, Chapman & Hall/CRC, 2002.