University of Cambridge

Part II of the Mathematical Tripos

Number Theory

Lectured by Jack Thorne, Michaelmas 2024–25

Notes by Avish Kumar
ak2461@cam.ac.uk
https://ak1089.github.io/maths/notes

Version 2.0

These notes are unofficial and may contain errors. While they are written and published with permission, they are not endorsed by the lecturer or University. Schedules for this course are subject to change in the future; these notes thus should not be relied on as a replacement for lectures in subsequent years.

Contents

1 Primes and Congruences

In this course, we study the ring of integers $\mathbb{Z}$. We especially will focus on primes, investigating questions like the limiting distribution of prime numbers.

1.1 Motivating Examples

For integers x, we define the prime counting function $\pi(x)$ to be the number of primes less than or equal to x, that is:

$$\pi(x) = \#\{p \mid 1 \leqslant p \leqslant x, p \text{ a prime}\}$$

The Riemann Hypothesis is equivalent to the proposition that $\forall x > 3, |\pi(x) - \mathrm{li}(x)| \leqslant \sqrt{x}\ln x$. What is this function $\mathrm{li}(x)$? It is the logarithmic integral:

$$\mathrm{li}(x) = \int_2^x \frac{dt}{\ln t}$$

We will see why this is true in due course.

We might also look at the study of Diophantine equations, which are equations where we desire integer solutions. A famous example is the equation underlying Fermat's last theorem:

$$X^N + Y^N = Z^N \text{ where } X, Y, Z, N \in \mathbb{Z},\ N \geqslant 3,\ XYZ \neq 0$$

which Fermat claimed had no solutions.

Thirdly, we might look at computational problems in prime factorisation: given some large N, can we "quickly" decide whether N is prime? If it is composite, can we "quickly" find its prime factorisation?

1.2 Prime Numbers

Of course, we all know what prime numbers are. Let's formalise some of their properties.

Proposition 1.1 (Division Algorithm)

If $a, b \in \mathbb{Z}$ with $b > 0$, there are unique $q, r \in \mathbb{Z}$ such that $qb + r = a$ and $0 \leqslant r < b$.

Proof: Take the set $\{a - nb : n \in \mathbb{Z}\}$. It has a least non-negative element: choose this and call it r. This is at least zero by construction, and we know $r < b$: if this were not the case, $r - b$ would also be in the set, also be non-negative, and be strictly smaller. Then q exists, since $r = a - nb$ for some n by definition.

This q must also be unique. If q and q' are different solutions, then $b(q - q') = r - r'$. The left side has magnitude at least b, but the right side has magnitude less than b: a contradiction! □

Definition 1.2 (Factor, Prime)

The integer a *divides* the integer b if there is an integer k such that $ka = b$. We write $a \mid b$, and say a is a *factor* of b, or that b is divisible by a. Otherwise, we write $a \nmid b$.

A *prime* number p is a positive integer with exactly two factors (1 and p). A number which is not prime is called *composite*. 1 is therefore *not* a prime number: it has only one factor, rather than exactly two.

Now, suppose we take some sequence of numbers $a_1, \ldots, a_n$, not all zero. Then, we take the set $I = \{\lambda_1 a_1 + \ldots + \lambda_n a_n\}$, where the λ_i are all in $\mathbb{Z}$. What can we say about this set? In fact, it must have a very particular structure: it is $d\mathbb{Z}$ for some d.

Proposition 1.3 (Highest Common Factor)

There exists some positive integer d such that this I is equal to the ring $d\mathbb{Z}$. This is called the highest common factor, or greatest common divisor.

Proof: Take the least positive element of I. Obviously, $d\mathbb{Z} \subseteq I$: d is a linear combination of the a_i, so linear multiples of it are also such linear combinations. Also, any k in I can be written as $k = qd + r$, where $0 \leqslant r < d$. But since d was minimal, r must be 0, so $k \in d\mathbb{Z}$, and so $I \subseteq d\mathbb{Z}$. □

Corollary: If e is a factor of every a_i, then it is also a factor of d.

Corollary: For integers a, b, c with a, b not both zero, if the highest common factor of a and b (written (a, b) for short) is a factor of c, then there are integers x and y such that $xa + yb = c$, and vice versa.

This corollary is known as *Bézout's identity.*

Remark 1.4 (Euclid's Algorithm)

Euclid's Algorithm allows us to compute the highest common factor of two numbers. If $a \geqslant b$:

$$\begin{aligned} a &= q_0 b + r_1 \\ b &= q_1 r_1 + r_2 \\ r_1 &= q_1 r_2 + r_3 \\ &\vdots \\ r_k &= q_{k+1} + r_{k+1} + 0 \end{aligned}$$

This always terminates in at most b steps, since $b, r_1, r_2, \ldots$ is a strictly decreasing sequence of integers. Thus we can find the highest common factor of two numbers in linear time.

Note that $(a, b) = (b, r_1) = (r_1, r_2) = \cdots = (r_{k+1}, 0) = r_{k+1}$.

We can also use this algorithm in reverse to find the x and y from Bézout's identity!

Now, let's think back to primes.

Proposition 1.5 (Prime Divisibility)

If $p \mid ab$, then $p \mid a$ or $p \mid b$ (or both).

Proof: Suppose $p \nmid a$. Then (a, p) is a positive factor of p, so is 1 or p. As $p \nmid a$, it cannot be p and so must be 1. So there are x, y such that $xa + yp = 1$, or multiplying, $xab + ypb = b$. Note that $p \mid ab \mid xab$, and $p \mid ypb$ (obviously), so p is a factor of the left hand side. But then $p \mid b$.

This argument in reverse shows the proposition. If $p \nmid b$, then $p \mid a$. So in fact one or the other must be true, as required. □

A prime factor of a number n is a factor $p \mid n$ which is prime. A *prime factorisation* of some number n is therefore a list of (not necessarily distinct primes) whose product is n. For example, $2 \times 2 \times 3 \times 5 = 60$ is a prime factorisation of 60. Then 2, 3, and 5 are prime factors of 60. In fact, these are the only prime factors, and we can prove this is true in general.

Theorem 1.6 (Fundamental Theorem of Arithmetic)

Every integer n can be expressed as a product of primes, unique up to reordering.

Proof: Existence is proved by strong induction. This is true for $n = 2$, as $2 \mid 2$. Suppose it is true for $n = 2, 3, 4, \ldots, n-1$. Then if n is prime, it is true for n, and if n is composite, then by definition n has some factor $1 < k < n$. This k has some prime factor by the inductive hypothesis, which is also a prime factor of n.

Uniqueness is proved similarly. Suppose

$$n = \prod_{i=1}^{k} p_i^{r_i} = \prod_{j=1}^{l} q_j^{s_j} \text{ with } p_i, q_j \text{ primes and } r_i, s_j \in \mathbb{N}$$

Then $p_1 \mid n$, so must divide some q_j. Divide by p_1 to get a strictly smaller expression. □

How do we find the prime factors p_i given a number N? Ideally, we have an algorithm much like Euclid's. First, we look at what we call polynomial-time algorithms.

Definition 1.7 (Polynomial-Time Algorithm)

An algorithm is said to run in *polynomial time* if there exist constants $b, c \in \mathbb{R}$ such that for all $N > 1$, the algorithm terminates after performing at most $b \times (\ln N)^c$ elementary operations. If the algorithm takes in multiple inputs, N refers to the maximum of the N_i.

Here, *elementary operations* are additions and multiplications of digits in a fixed base.

Note: This is the class of "fast" algorithms. Exponential-time algorithms exist, and cannot be bounded by this expression, so they can often take much longer to run.

Note: This is primarily an asymptotic property. It is possible that a polynomial-time algorithm has extraordinarily large bounding constants b and c, such that an exponential-time algorithm can outperform it on most reasonably-sized inputs. In the long run, rhough, the former algorithm will dominate, as exponentials outgrow polynomials asymptotically.

Euclid's algorithm is polynomial-time. So is primality testing: this was proved in 2002. The naïve factorisation algorithm of testing division up to $\sqrt{N}$ is *not* polynomial-time: asymptotically, this is larger than any power of $\ln N$. This is not the best algorithm, but we currently do not know of any polynomial-time factorisation algorithms.

Theorem 1.8 (Infinitude of Primes)

There are infinitely many prime numbers: $\pi(x)$ is unbounded.

Proof: Suppose not, and there are finitely many prime numbers. Take the product of all of them: $N = 2 \times 3 \times 5 \times \ldots \times p_{\text{largest}}$. Then $N + 1$ is not divisible by any of these primes. However, every number has a prime factor, so this is a contradiction. □

In fact, the best way to find large primes (say, on the order of 50 digits), is to generate numbers of the right size at random and apply a fast primality test! How long this takes depends on the density of prime numbers, which depends on the behaviour of $\pi(x)$.

Another way to find primes is to look at certain patterns, such as that of the *Mersenne primes*. If p is a prime, then $2^p - 1$ is often prime, and more importantly there is a very fast test to see if it is. The largest known prime known as of the end of 2024 is $2^{82589933} - 1$, and it was found while this course was being lectured!

1.3 Modular Arithmetic

Modular arithmetic will be a large focus of this section. We now define it formally.

Definition 1.9 (Congruence)

Fix $N \in \mathbb{N}$. If $a, b \in \mathbb{Z}$, we write

$$a \equiv b \pmod{N} \iff N \mid (a-b)$$

and say that a is congruent to b modulo N. We write $\mathbb{Z}/N\mathbb{Z}$ for the quotient ring of $\mathbb{Z}$ under the ideal $N\mathbb{Z}$. Note that this is an equivalence relation on $\mathbb{Z}$ with classes $a + N\mathbb{Z}$.

Note: Addition and multiplication are well-defined modulo N: $(a+N\mathbb{Z})+(b+N\mathbb{Z}) = (a+b)+N\mathbb{Z}$ and $(a+N\mathbb{Z})(b+N\mathbb{Z}) = ab+N\mathbb{Z}$.

Proposition 1.10 (Units modulo N)

Let $a \in \mathbb{Z}$. Then the following are equivalent:

(a) $\gcd(a, N) = 1$

(b) $\exists b \in \mathbb{Z}$ s.t. $ab \equiv 1 \pmod{N}$

(c) $a + N\mathbb{Z}$ generates the group $(\mathbb{Z}/N\mathbb{Z}, +)$.

Proof: (a $\Leftrightarrow$ b) $\gcd(a, N) = 1 \iff \exists b, y \in \mathbb{Z}$ s.t. $ab + yN = 1 \iff ab \equiv 1 \pmod{N}$.

(b $\Rightarrow$ c) $1 + N\mathbb{Z}$ obviously generates the group: any b can be generated with b additions. We know a generates 1, so must generate the whole group.

(c $\Rightarrow$ b) If $a + N\mathbb{Z}$ is a generator, then $\exists b \in N$ s.t. $ab + N\mathbb{Z} = 1 + N\mathbb{Z}$, so we are done. □

We write $(\mathbb{Z}/N\mathbb{Z})^\times \subseteq \mathbb{Z}/N\mathbb{Z}$ for the set of $a + N\mathbb{Z}$ satisfying the previous proposition, and often identify it as a multiplicative group. We write $\phi(N)$ for the cardinality of this set, which is

$$\phi(N) = \#\{1 \leqslant a \leqslant N \text{ with } \gcd(a, N) = 1\}.$$

This is called *Euler's totient function.*

Corollary: $(\mathbb{Z}/N\mathbb{Z})^\times$ is a group under multiplication.

Corollary: If $N > 1$, then $\phi(N) \leqslant N - 1$, with equality if and only if N is prime.

Corollary: The cyclic group C_N of order N has precisely $\phi(N)$ elements with order N exactly.

Theorem 1.11 (Euler-Fermat Theorem)

Suppose $a, N \in \mathbb{Z}, N > 1, \gcd(a, N) = 1$. Then $a^{\phi(N)} \equiv 1 \pmod{N}$.

Proof: Observe that $(\mathbb{Z}/N\mathbb{Z})^\times$ is a group of order $\phi(N)$. Then, by Lagrange's theorem, we have that $(a + N\mathbb{Z})^{\phi(N)} = a^{\phi(N)} + N\mathbb{Z} = 1 + N\mathbb{Z}$. □

Theorem 1.12 (Fermat's Little Theorem)

For any prime p and integer a, $a^p \equiv a \pmod{p}$.

Proof: If $p \mid a$, then $a^p \equiv a \equiv 0 \pmod{p}$.

Otherwise, $\gcd(a, p) = 1$. Then $a^{\phi(p)} = a^{p-1} \equiv 1 \pmod{p}$. Multiplying by a yields $a^p \equiv a \pmod{p}$ exactly as required. □

Example 1.13 (Simultaneous Congruences)

Can we find $x \in \mathbb{Z}$ such that $x \equiv 3 \pmod{10}$ and $x \equiv 7 \pmod{13}$?

We can obviously do this if we find u and v such that

$$\begin{aligned} u &\equiv 1 \pmod{10} & u &\equiv 0 \pmod{13} \\ v &\equiv 0 \pmod{10} & v &\equiv 1 \pmod{13} \end{aligned}$$

(by taking $x = 3u + 7v$). By Euclid, we can find a, b such that $10a + 13b = \gcd(10, 13) = 1$. In fact, $a = -9$ and $b = 7$ works. Then, we take $u = 13b$ and $v = 10a$.

So finally, $x = 39b + 70a$, which is -357. Indeed, this is a solution! We can also add multiples of $10 \times 13 = 130$, to get eg. 33 as a positive solution.

Note: From now on, we write (a, b) for $\gcd(a, b)$. If $(a, b) = 1$, we say that a and b are *coprime*.

Theorem 1.14 (Chinese Remainder Theorem)

Suppose we are given integers $m_1 \dots m_k$ satisfying $\forall i, m_i > 1$ and $\forall i, j$ we have $(m_i, m_j) = 1$. Then for given integers $a_1 \dots a_k$, the simultaneous congruence

$$x \equiv \begin{cases} a_1 \pmod{m_i} \\ \vdots \\ a_k \pmod{m_k} \end{cases}$$

has a solution which is unique modulo $M = \prod m_i$.

Proof: (Uniqueness) If x, y are two solutions, then $x \equiv y \pmod{m_i}$ for all i. Then $m_i \mid (x - y)$ for all i, and as they are pairwise coprime, we have $M \mid (x - y)$.

(Existence) Define $M_i = \prod_{j \neq i} m_j$. Then $(m_i, M_i) = 1$. By Bézout's identity, there are then integers such that $x_i m_i + y_i M_i = 1$.

In particular, $y_i M_i = 1 \pmod{m_i}$, and $y_i M_i = 1 \pmod{m_j}$ for all $j \neq i$. Thus

$$x = \sum_{i=1}^{k} a_i y_i M_i$$

is a solution, which proves the theorem. □

Theorem 1.15 (Ring Isomorphism)

Given moduli as before, the function

$$\begin{aligned} \theta : \mathbb{Z}/M\mathbb{Z} &\to \mathbb{Z}/m_1\mathbb{Z} \times \cdots \times \mathbb{Z}/m_k\mathbb{Z} \\ a + M\mathbb{Z} &\mapsto (a + m_1\mathbb{Z}, \dots, a + m_k\mathbb{Z}) \end{aligned}$$

is a ring isomorphism: it is bijective and respects addition and multiplication.

Proof: (Bijection) True by the Chinese Remainder Theorem (1.14).

(Homomorphism) The codomain ring is defined componentwise, so we need only check that the map onto $\mathbb{Z}/m_i\mathbb{Z}$ respects addition and multiplication.

But this follows immediately from the definition of these operations! So this function θ really is a ring isomorphism. Thus the product of the individual rings is isomorphic to the main ring. □

Corollary: θ restricts to a group isomorphism $(\mathbb{Z}/M\mathbb{Z})^\times \to (\mathbb{Z}/m_1\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/m_k\mathbb{Z})^\times$.

Proof: θ gives you a bijection between the multiplicative groups $(\mathbb{Z}/M\mathbb{Z})^\times$ and the elements of the product ring which have a multiplicative inverse. But this target ring is defined componentwise, which is the product ring of elements which multiplicative inverses in each component ring! This is simply $(\mathbb{Z}/m_i\mathbb{Z})^\times$ for each i, as desired. □

We will soon show that if N is a prime, then $(\mathbb{Z}/N\mathbb{Z})^\times$ is cyclic. In fact, with this corollary, we will show that if $N > 1$ is an odd squarefree integer, then N is a prime *only* if $(\mathbb{Z}/N\mathbb{Z})^\times$ is cyclic.

Definition 1.16 (Multiplicative Functions)

Let $f : \mathbb{N} \to \mathbb{C}$ be a function. We say f is multiplicative if

$$\forall m, n \in \mathbb{N} \text{ s.t. } (m, n) = 1, f(mn) = f(m)f(n)$$

and we say f is *totally* multiplicative if we can drop the restriction $(m, n) = 1$. These are genuinely different definitions, as we shall soon see.

Corollary: Clearly, the constant function $f(n) = 1$ and the identity function $f(n) = n$ are both totally multiplicative.

Corollary: The totient function ϕ is *not*, as $\phi(2) = 1$ but $\phi(2 \cdot 2) = 2 \neq 1 \cdot 1$.

However, we now show that the totient function is multiplicative, proving that our two definitions are in fact distinct.

Proposition 1.17 (Multiplicative Totient Function)

The totient function ϕ is multiplicative.

Proof: $\phi(m) = \#(\mathbb{Z}/m\mathbb{Z})^\times$. We need to show that for all coprime integers m and n,

$$\#(\mathbb{Z}/mn\mathbb{Z})^\times = \#(\mathbb{Z}/m\mathbb{Z})^\times \cdot \#(\mathbb{Z}/n\mathbb{Z})^\times.$$

But this is true by the fact that the former group is isomorphic to the product of the other two. □

Proposition 1.18 (Construction of Multiplicative Functions)

Let f be a multiplicative function. Define $g : \mathbb{N} \to \mathbb{C}$ by $g(n) = \sum_{d|n} f(d)$ (summing over all factors of n). Then g is also multiplicative.

Proof: Take coprime m, n. We must show $g(mn) = g(m)g(n)$. Then

$$g(mn) = \sum_{d|mn} f(d) = \sum_{d_1|m, d_2|n} f(d_1 d_2) = \sum_{d_1|m, d_2|n} f(d_1)f(d_2) = \left(\sum_{d_1|m} f(d_1)\right)\left(\sum_{d_2|n} f(d_2)\right)$$

but this is exactly $g(m)g(n)$, as required. □

Proposition 1.19 (Properties of ϕ)

If p is prime, and $k \in \mathbb{N}$, then $\phi(p^k) = p^k - p^{k-1}$.

For any $N \in \mathbb{N}$, $\phi(N) = N \prod_{p \text{ prime } | N} \left(1 - \frac{1}{p}\right)$.

Also, for any $N \in \mathbb{N}$, $\sum_{d|N} \phi(d) = N$.

Proof: Firstly, $\phi(p^k) = \#\left\{1 \leqslant a \leqslant p^k \text{ with } (a, p^k) = 1\right\}$. For all members of this set, $p \nmid a$. So this is the set of nonmultiples of p, of which there are p^{k-1}.

Secondly, we can write $N = \prod_{i=1}^{r} p_i^{k_i}$, where distinct.

Thirdly, define $g(n) = \sum_{d|n} \phi(d)$. This is multiplicative. We want to show that $g(n) = n$ for all n. Since both sides are multiplicative, it's enough to check this when n is a prime power. We can check this easily:

$$g(p^k) = \sum_{d|p^k} \phi(d) = \sum_{i=0}^{k} \phi(p^i) = 1 + \sum_{i=1}^{k} (p^i - p^{i-1}) = p^k$$

which means this property holds for all prime powers, and thus for all n. □

1.4 Polynomials and Cyclic Groups

If $n \in \mathbb{N}$, then a *polynomial* over $\mathbb{Z}/N\mathbb{Z}$ is an expression $f(X) = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_0$, where $a_i \in \mathbb{Z}/N\mathbb{Z}$. These can be summed and multiplied as usual. We write $\mathbb{Z}/N\mathbb{Z}[X]$ for the set of such polynomials in X: note that this is also a ring.

If $f(X)$ is such a polynomial, and $x \in \mathbb{Z}/N\mathbb{Z}$, then we write $f(a) = a_n x^n + \cdots + a_0$, which is just a sum of integers in this ring. We say the *solutions* to the equation $f(X) = 0$ in the ring are the elements such that $f(a) \equiv 0 \pmod{N}$.

Example 1.20 (Solutions to Polynomials)

The equation $X^2 + 2 = 0$ in $\mathbb{Z}/5\mathbb{Z}$ has no solutions. (Proof by exhaustion).

The equation $X^3 + 1 = 0$ in $\mathbb{Z}/7\mathbb{Z}$ has the solutions $\{3, 5, 6\}$.

The equation $X^2 - 1 = 0$ in $\mathbb{Z}/8\mathbb{Z}$ has the solutions $\{1, 3, 5, 7\}$.

In this last case, note that the polynomial has degree 2, but four solutions. In fact, this is only possible because 8 is not prime.

Theorem 1.21 (Lagrange's Theorem)

Let p be a prime, and $f(x) \in \mathbb{Z}/p\mathbb{Z}[X]$ be a polynomial of degree n with $a_n \not\equiv 0 \pmod{p}$. Then $f(X) = 0$ has at most n solutions.

Proof: We use induction on the degree of the polynomial n. Certainly, the base case $n = 0$ works: if $f(X) = a_0 \not\equiv 0 \pmod{p}$, then there are no solutions.

Now suppose $n > 0$. If there are no solutions, then we are done. Now suppose there is a solution a. Then note that for all $j \geqslant 1$, we have $X^j - a^j = (X - a)(X^{j-1} + aX^{j-2} + \cdots + a^j - 1)$. So $f(X) = f(X) - f(a) = (X - a)g(X)$, where g is another polynomial with leading term $a_n X^{n-1}$.

Suppose b is a solution of $f(X) = 0$. Then $0 \equiv f(b) \equiv (b - a)g(b) \pmod{p}$. So either $(b - a) \equiv 0 \pmod{p}$, or not, in which case $g(b) \equiv 0 \pmod{p}$. This step uses the primality of p.

If so, then b would also be a solution of g. By induction, there are at most $n - 1$ of these. So there are at most n total, now including a. □

Theorem 1.22 (Prime Cyclic Groups)

Let p be a prime. Then $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic.

Proof: Let $G = (\mathbb{Z}/p\mathbb{Z})^\times$. Then $|G| = p-1$. We know that

$$\sum_{d|p-1} \phi(d) = p-1 \quad \text{and} \quad \text{ord}(g) \mid p-1 \forall g \in G$$

Then $p-1$ is equal to the sum of N_d over divisors d of $p-1$, where N_d is the number of elements in G with order exactly d.

G is cyclic if and only if N_{p-1} is non-zero. Suppose for a contradiction that this is not true. Then

$$\sum_{d|p-1} \phi(d) = \sum_{d|p-1, d\neq p-1} N_d$$

so we must have $N_d > \phi(d)$ for some $d \mid p-1$.

Take this d. Then take some element $a \in G$ of order d, and consider the subgroup $\langle a \rangle$ generated by a: it is cyclic and has d elements.

We have seen previously that any cyclic group of d elements has $\phi(d)$ elements of order d. So there must exist some element $b \in G$ of order d that is not in this group.

The elements of $\langle a \rangle$ have order dividing d, so they are all solutions to the polynomial $X^d - 1 = 0$ in $\mathbb{Z}/p\mathbb{Z}$. b is also a solution to this polynomial congruence.

But then there would be $d+1$ distinct solutions to a polynomial equation of degree d. This would contradict Lagrange's Theorem (1.21), and therefore $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic. □

Note: This group is called the multiplicative group of the finite field of $\mathbb{Z}$ modulo p. It is a cyclic group, and it has $p-1$ elements, so in fact it is isomorphic to C_{p-1}.

Definition 1.23 (Primitive Root)

If p is a prime, and a is an integer, we say a is a *primitive root* modulo p if $(a,p) = 1$ and $a \pmod p$ generates the group $(\mathbb{Z}/p\mathbb{Z})^\times$.

How can we find such primitive roots?

Example 1.24 (Primitive Root)

Take $p = 7$. Then for example:

$$2 \to 4 \to 1 \to 2 \to 4 \to 1 \text{ is not a primitive root.}$$
$$3 \to 2 \to 6 \to 4 \to 5 \to 1 \text{ is a primitive root!}$$

In fact, the primitive roots modulo 7 are only 3 and 5.

Now, take $p = 19$. We want to know whether $a = 2$ is a primitive root modulo 19. We set d as the order of 2 mod 19 in the group $(\mathbb{Z}/19\mathbb{Z})^\times$. We know that d divides the order of the group, which is 18. Note that $d = 18 \iff 2$ is a primitive root mod 19.

$$d = 18 \iff d \nmid 9 \wedge d \nmid 6 \iff 2^9 \not\equiv 1 \wedge 2^6 \not\equiv 1 \pmod{19}.$$

Then, we can check $2^6 = 64 \equiv 7 \pmod{19}$, so $2^9 \equiv 56 \equiv 18 \pmod{19}$.

This means 2 is a primitive root modulo 19.

Corollary: In general, for primes p, if $a \in \mathbb{Z}$ is such that $(a,p) = 1$, then a is a primitive root if and only if $a^{(p-1)/q} \not\equiv 1 \pmod p$ for every prime $q \mid p-1$.

Remark 1.25 (Difficulty of Finding Primitive Roots)

In general, carrying out this test is hard: it requires knowledge of the prime factorisation of $p-1$. There is no known polynomial-time algorithm to find a primitive root modulo p.

However, if the Generalised Riemann Hypothesis is true, then

$$\exists c > 0 \text{ s.t. } \forall \text{ primes } p, \exists a : 1 < a < c \ln(p)^6$$

where a here is a primitive root modulo p.

Now, we consider $(\mathbb{Z}/p^k\mathbb{Z})^\times$, where p is a prime and $k \in \mathbb{N}$.

Proposition 1.26 (Prime Power Generators)

Let p be an odd prime, $k \in \mathbb{N}$, and $x, y \in \mathbb{Z}$. Then

$$x \equiv 1 + p^k y \pmod{p^{k+1}} \implies x^p \equiv 1 + p^{k+1} y \pmod{p^{k+2}}.$$

Moreover, we have:

$$(1+py)^{p^k} \equiv 1 + p^{k+1} y \pmod{p^{k+2}}.$$

Proof: Firstly, write $x = 1 + p^k y + p^{k+1} z$ for some z. Then

$$x^p = \left(1 + p^k y\right)^p + \sum_{j=1}^{p} \binom{p}{j} \left(1 + p^k y\right)^{p-j} \left(p^{k+1} z\right)^j$$

Each summand is then divisible by p^{k+2}. If $0 < j < p$, then

$$p \mid \binom{p}{j} \implies p^{k+2} \mid \binom{p}{j} p^{k+1}$$

Also, $p^{k+2} \mid p^{p(k+1)} \mid$ the summand for which $j = p$.

So we can assume that $x = 1 + p^k y$. Then the sum becomes

$$x^p = \left(1 + p^k y\right)^p = 1 + p^{k+1} y + \sum_{j=2}^{p} \binom{p}{j} \left(p^k y\right)^j$$

Each summand is still divisible by p^{k+2}, so the result holds, given $p \geqslant 3$.

We can apply this part k times to get the second result. □

Theorem 1.27 (Cyclic Groups)

If p is an odd prime, then $(\mathbb{Z}/p^k\mathbb{Z})^\times$ is cyclic.

Proof: Assume $k \geqslant 2$. Then $\#(\mathbb{Z}/p^k\mathbb{Z})^\times = \phi(p^k) = p^k - p^{k-1} = p^{k-1}(p-1)$.

Let d be the order of a mod p^k. Then $d \mid (p-1)p^{k-1}$, so we must show it is equal to this quantity.

Note that there is a surjective homomorphism $(\mathbb{Z}/p^k\mathbb{Z})^\times \to (\mathbb{Z}/p\mathbb{Z})^\times$ which sends $b + p^k\mathbb{Z}$ to $b + p\mathbb{Z}$. The image of $a + p^k\mathbb{Z}$ under this homomorphism is $p-1$ by assumption. Then $(p-1) \mid d$, so $d = (p-1)p^j$ for some $0 \leqslant j < k$.

Now let $x = a^{p-1} = 1 + py$ for some $y \in \mathbb{Z}$ coprime to p. The order of x mod p^k in the group $(\mathbb{Z}/p^k\mathbb{Z})^\times$ is p^j. We must show that x has order p^{k-1} modulo p^k, or equivalently that $x^{p^{k-2}} \not\equiv 1 \pmod{p^k}$.

If $k = 2$, we want $x \not\equiv 1 \pmod{p^2}$, which is true by assumption. So take $k \geqslant 3$, ie. $k - 2 \geqslant 1$.

Then by Proposition 1.26, we know that

$$x^{p^{k-2}} = (1 + py)^{p^{k-2}} \equiv 1 + p^{k-1}y \not\equiv 1 \pmod{p}$$

Let $b \in \mathbb{Z}$ be a primitive root modulo p. If $b^{p-1} \not\equiv 1 \pmod{p^2}$, we're done, so assume that it is. Take $a = (1 + p)b$. Then $a \equiv b \pmod{p}$, so a is a primitive root mod p. We have

$$a^{p-1} = (1 + p)^{p-1}b^{p-1} \equiv 1 + p(p - 1) \equiv 1 - p \not\equiv 1 \pmod{p^2}$$

□

Example 1.28 (Primitive Roots and Generators)

We saw that 3 is a primitive root modulo 7. Does it generate $(\mathbb{Z}/7^k\mathbb{Z})^\times$ for all $k \geqslant 1$?

This holds if $3^6 \not\equiv 1 \pmod{49}$. In fact, $3^6 = 729 \equiv 43 \pmod{49}$. So in fact it does generate all of these groups!

Remark 1.29 (Why "Odd" Prime?)

Many statements in this section have referred to p being an odd prime. In fact, this is for a good reason: not every result carries over for $p = 2$, the only even prime. For example:

$$(1 + py)^{p^k} \equiv 1 + p^{k+1}y \pmod{p^{k+2}}$$

is actually false when $p = 2$ and $k = 1$, as $9 \not\equiv 4 \pmod{8}$.

However, it does hold when $p = 2$ and $k \geqslant 2$. The group

$$\left\{x + 2^k\mathbb{Z} \in (\mathbb{Z}/2^k\mathbb{Z})^\times : x \equiv 1 \pmod{4}\right\}$$

is cyclic, and is generated by $5 + 2^k\mathbb{Z}$.

2 Quadratic Reciprocity

Having studied the prime numbers, we move on to *quadratic reciprocity*, first introducing a new way in which to analyse squares modulo p.

Definition 2.1 (Quadratic Residue)

Let p be a prime and a an integer such that $(a, p) = 1$. Then we say that a is a *quadratic residue* modulo p if

$$\exists\, x \in \mathbb{Z}/p\mathbb{Z} \text{ s.t. } x^2 - a = 0,$$

and a *quadratic non-residue* otherwise. Equivalently, a is a quadratic residue mod p if and only if $a + p\mathbb{Z}$ is a square in $(\mathbb{Z}/p\mathbb{Z})^\times$.

Note: Suppose $p = 7$. Then, since $1^2 \equiv 6^2 \equiv 1$, $2^2 \equiv 5^2 \equiv 4$, and $3^2 \equiv 4^2 \equiv 2$ modulo 7, the quadratic residues modulo 7 are precisely 1, 2, and 4.

Proposition 2.2 (Number of Quadratic Residues)

If p is an odd prime, then there are precisely $\frac{p-1}{2}$ quadratic residues modulo p.

Proof: Consider the map $\sigma : (\mathbb{Z}/p\mathbb{Z})^\times \to (\mathbb{Z}/p\mathbb{Z})^\times, x \mapsto x^2$. We need to show that the image of σ has $\frac{p-1}{2}$ elements, so it suffices to show that the preimage of each class has exactly two elements.

if $x, y \in (\mathbb{Z}/p\mathbb{Z})^\times$ and $x^2 \equiv y^2 \pmod{y}$, then $(x - y)(x + y) \equiv 0 \pmod{p}$. Thus $x \equiv \pm y \pmod{p}$, so the preimage of x^2 has precisely the two elements $\{x, -x\}$. □

Definition 2.3 (Legendre Symbol)

For p an odd prime and $a \in \mathbb{Z}$, we write

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

Proposition 2.4 (Euler's Criterion)

If p is an odd prime, then

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}.$$

Proof: If $p \mid a$ then both sides are 0 modulo p. So assume $p \nmid a$. If $a \equiv x^2 \pmod{p}$, then

$$\left(\frac{a}{p}\right) = 1 \text{ and } a^{\frac{p-1}{2}} \equiv x^{p-1} \equiv 1 \pmod{p}$$

by the Euler-Fermat Theorem (1.11). If a is a quadratic non-residue, then

$$\left(\frac{a}{p}\right) = -1 \text{ and } \left(a^{\frac{p-1}{2}}\right)^2 \equiv a^{p-1} \equiv 1 \pmod{p}.$$

Therefore, $\left(a^{\frac{p-1}{2}}\right) \equiv \pm 1 \pmod{p}$. We need to show it is not $+1$.

By Lagrange's Theorem (1.21) know $x^{\frac{p-1}{2}} = 0$ has at most $\frac{p-1}{2}$ solutions in $\mathbb{Z}/p\mathbb{Z}$. But we also know there are at least this many solutions, given by the quadratic residues. □

Corollary: For odd primes p,

$$\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}} = \begin{cases} +1 & p \equiv 1 \pmod 4 \\ -1 & p \equiv 3 \pmod 4 \end{cases}$$

Definition 2.5 (Closest Integer to Zero)

Suppose p is an odd prime and $a \in Z$. Then there is a unique integer $b \in a + p\mathbb{Z}$ such that $-p/2 < b < p/2$.

We write $\langle a \rangle = b$ for this integer.

Theorem 2.6 (Gauss's Lemma)

Let p be an odd prime and a a coprime integer. Then

$$\left(\frac{a}{p}\right) = (-1)^{\mu} \text{ where } \mu = \#\left\{j \in \mathbb{Z} : 0 < j < p/2, \langle ja \rangle < 0\right\}$$

Proof: Consider the expressions

$$\left(\frac{p-1}{2}\right)! \equiv \prod_{j=1}^{\frac{p-1}{2}} j \quad \text{and} \quad \left(\frac{p-1}{2}\right)! \equiv \prod_{j=1}^{\frac{p-1}{2}} aj.$$

These represent each side of the inequality respectively. □

Definition 2.7 (Floor)

For $x \in \mathbb{R}$, the *floor* of x is defined as

$$\lfloor x \rfloor = \sup\left\{n \in \mathbb{Z} : n \leqslant x\right\}.$$

Note: For $x \in \mathbb{Z}$, $\lfloor x \rfloor = x$. Also, for all $x \in \mathbb{R}$, we have $x - 1 < \lfloor x \rfloor \leqslant x$.

Example 2.8 (Evaluating Legendre Symbols)

Let's try and evaluate one of these expressions. We know that

$$\left(\frac{3}{p}\right) = (-1)^{\mu} \text{ where } \mu = \#\left\{j \in \mathbb{Z} : 0 < j < \frac{p}{2}, \langle 3j \rangle < 0\right\}$$

We can assume $p > 3$, since the other cases are trivial to compute.

If $0 < j < p/6$, then $0 < 3j < p/2$, so $\langle 3j \rangle > 0$.

If $p/6 < j < 2p/6$, then $p/2 < 3j < p$, so $\langle 3j \rangle < 0$.

Finally, if $2p/6 < j < 3p/6$, then $p < 3j < 3p/2$, so $\langle 3j \rangle > 0$.

So only the second case contributes! Thus we can write

$$\left(\frac{3}{p}\right) = (-1)^{\mu} \text{ where } \mu = \#\left\{j \in \mathbb{Z} : \frac{p}{6} < j < \frac{p}{3}\right\} = \left\lfloor \frac{p}{3} \right\rfloor - \left\lfloor \frac{p}{6} \right\rfloor.$$

So this is the value of the Legendre symbol in closed form!

Suppose we take $a \in \mathbb{Z}$, $p \nmid a$ a prime. Then by definition, $\langle aj \rangle = aj - pc$, where c is the unique integer such that $-p/2 < aj - pc < p/2$. We can then express μ as the number of elements in:

$$\left\{(b,c) \in \mathbb{Z}^2 : 0 < b < p/2,\ -p/2 < ab - pc < 0\right\}.$$

Theorem 2.9 (Law of Quadratic Reciprocity)

Suppose p and q are distinct odd primes. Then

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

Equivalently, we can express them in terms of each other:

$$\left(\frac{p}{q}\right) = \begin{cases} -\left(\frac{q}{p}\right) & p \equiv q \equiv 3 \pmod 4 \\ \left(\frac{q}{p}\right) & \text{otherwise} \end{cases}$$

Proof: By Gauss's Lemma, we have

$$\left(\frac{q}{p}\right) = (-1)^{|A|} \text{ where } A = \left\{(b,c) \in \mathbb{Z}^2 : 0 < b < \frac{p}{2}, -\frac{p}{2} < qb - pc < 0\right\}$$

Similarly, we have

$$\begin{aligned}\left(\frac{p}{q}\right) &= (-1)^{|B|} \text{ where } B = \left\{(b,c) \in \mathbb{Z}^2 : 0 < b < \frac{q}{2}, -\frac{q}{2} < pb - qc < 0\right\} \\ &= (-1)^{|C|} \text{ where } C = \left\{(b,c) \in \mathbb{Z}^2 : 0 < c < \frac{q}{2}, 0 < qb - pc < \frac{q}{2}\right\}\end{aligned}$$

by renaming. Now let $S = \left\{(b,c) \in \mathbb{Z}^2 : 0 < b < p/2, 0 < c < q/2\right\}$. Then the size of S is precisely the exponent in the right hand size of the equality we want to demonstrate.

$$|S| = \frac{p-1}{2} \times \frac{q-1}{2}.$$

We claim that A and C are disjoint subsets of S.

If the tuple $(b,c) \in A$, then $0 < b < p/2$, and $pc > qb$, so $c > qb/p > 0$. Moreover, $pc < qb + (p/2)$, so $c < (qb/p) + (1/2) < (q+1)/2$. Since $c \in \mathbb{Z}$, $c < q/2$, so $(b,c) \in S$, ie. $A \subseteq C$.

By the same argument, $C \subseteq S$. We now need to show them to be disjoint. This is clear, as $qb - pc < 0$ within A but is positive within C.

Now, we must show that $(-1)^{|A|+|C|} = (-1)^{|S|}$. We will show that $|S \setminus (A \cup C)|$ is even. Take

$$X = \{(b,c) \in S : qb - pc < -p/2\} \qquad Y = \{(b,c) \in S : qb - pc > q/2\}$$

Note that A, C, X, and Y are pairwise disjoint, and that $S \setminus (A \cup C) = X \cup Y$ (the four cover the set). We aim to show that $|X \cup Y|$ is even, so it suffices to show $|X| = |Y|$.

Let $f : S \to S$ be the function

$$(b,c) \mapsto \left(\frac{p+1}{2} - b, \frac{q+1}{2} - c\right)$$

This is a bijection. We will show that $f(X) \subseteq Y$ and $f(Y) \subseteq X$.

Suppose $(b,c) \in X$. Then $qb - pc < -p/2$, and $-qb + pc > p/2$. Then

$$q\left(\frac{p+1}{2} - b\right) - p\left(\frac{q+1}{2} - c\right) = \frac{q}{2} - qb - \frac{p}{2} + pc > \frac{q}{2} \implies f(b,c) \in Y.$$

A similar argument holds to show $(b,c) \in Y \implies f(b,c) \in X$.

Thus $|X| = |Y|$, so the law holds. □

Example 2.10 (Evaluating Legendre Symbols again)

Let's now evaluate the Legendre symbol

$$\left(\frac{3}{p}\right) \text{ where } p > 3 \text{ is a prime.}$$

By the Law of Quadratic Reciprocity (Theorem 2.9), we have

$$\left(\frac{3}{p}\right) = \begin{cases} \left(\frac{p}{3}\right) & p \equiv 1 \pmod 4 \\ -\left(\frac{p}{3}\right) & p \equiv 3 \pmod 4 \end{cases}$$

From the previous example (2.8), we know that

$$p \equiv 1 \pmod 3 \implies \left(\frac{p}{3}\right) = 1 \qquad p \equiv 2 \pmod 3 \implies \left(\frac{p}{3}\right) = -1$$

By the Chinese Remainder Theorem (1.14), we can convert these congruences modulo 4 and 3 into a single congruence modulo 12, which is

$$\left(\frac{3}{p}\right) = \begin{cases} +1 & p \equiv \pm 1 \pmod{12} \\ -1 & p \equiv \pm 5 \pmod{12} \end{cases}$$

This is a closed-form solution to the Legendre symbol!

Example 2.11 (Quadratic Solutions)

Does $X^2 - 19 = 0$ have a solution in $\mathbb{Z}/73\mathbb{Z}$? Well, since 73 is prime:

$$\left(\frac{19}{73}\right) = \left(\frac{73}{19}\right) = \left(\frac{16}{19}\right) = 1 \text{ (as } 16 = 4^2\text{)}.$$

So there is a solution, as this is the definition of the Legendre symbol being 1.

In fact, $26^2 = 676 = 9 \times 73 + 19$, and $47^2 = 2209 = 30 \times 73 + 19$, so these are our solutions. They are the *only* two solutions, and $26 + 47 = 73$.

Example 2.12 (Computing Large Legendre Symbols)

Given that 7411 and 9283 are prime and both congruent to 3 modulo 4,

$$\left(\frac{7411}{9283}\right) = -\left(\frac{9283}{7411}\right) = -\left(\frac{1872}{7411}\right)$$

We have $1872 = 2^4 \times 3^2 \times 13$, so this is equal to

$$-\left(\frac{1872}{7411}\right) = \left(\frac{13}{7411}\right) = -\left(\frac{1}{13}\right) = -1.$$

So even moderately large numbers lend themselves to quick computing, as long as we can use their factorisation to simplify the appropriate Legendre symbol.

What if we don't have a nice factorisation of our number on hand? It would still be convenient to compute Legendre symbols easily.

To accomplish this, we need to extend the definition of the symbol.

2.1 Jacobi Symbols

We now define a variation of the Legendre symbol, which holds even when p is not a prime.

Definition 2.13 (Jacobi Symbol)

Let $N \in \mathbb{N}$ be odd, with $a \in \mathbb{Z}$. Then we define the Jacobi symbol

$$N = \prod_{i=1}^{k} p_i^{r_i} \implies \left(\frac{a}{N}\right)_{\text{Jacobi}} = \prod_{i=1}^{k} \left(\frac{a}{p_i}\right)_{\text{Legendre}}^{r_i}$$

These agree when N is a prime. Note that when N is not a prime, the Jacobi symbol does *not* tell you whether a has a square in $\mathbb{Z}/N\mathbb{Z}$.

Note: We sometimes write (a/N) for the Jacobi symbol. Since division is not really considered when working in the ring of integers, this is not ambiguous.

If $N = 1$, then $(a/N) = 1$. If $(a, N) > 1$, then $(a/N) = 0$, as there is a p dividing a and N.

Generally, if $N = pq$, then $(a, N) = 1 \implies a \pmod N \in (\mathbb{Z}/N\mathbb{Z})^\times \cong (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times$. So squares modulo N are also squares modulo p and q, by the Chinese Remainder Theorem (1.14). Equivalently, $a \bmod N = pq$ is a square if and only if $(a/p) = (a/q) = 1$.

Now, consider that the product of these two symbols is equal to (a/N). If this is 1, then either both are 1 (and the condition is satisfied), or both are -1.

Corollary: $(a/N) = 1$ is necessary but insufficient to ensure that a is a square mod N.

Proposition 2.14 (Jacobi Multiplicity)

Let M and N be odd, with a and b integers. Then

1. $a \equiv b \pmod N \implies (a/N) = (b/N)$. Only the residue modulo N matters.
2. $(ab/N) = (a/N) \cdot (b/N)$. That is, the first argument is multiplicative.
3. $(a/MN) = (a/M) \cdot (a/N)$. That is, the second argument is also multiplicative.

Proof: It is fairly simple to show all of these properties.

1. If $a \equiv b \pmod N$, then $a \equiv b \pmod{p_j}$ for all $p_j \mid N$. This means $(a/p_j) = (b/p_j)$ for all j. Since these symbols only depend on the congruence class of the top modulo the bottom, they are the same, so the result holds.
2. This follows from the definition, writing N out as a product of primes.
3. This follows from the definition, writing N and M out as a product of primes.

So all three of these properties carry over from Legendre symbols. □

Proposition 2.15 (Jacobi Symbols)

If N is odd, then

$$\left(\frac{-1}{N}\right) = (-1)^{\frac{N-1}{2}} \quad \text{and} \quad \left(\frac{2}{N}\right) = (-1)^{\frac{N^2-1}{8}}$$

Proof: It's easy to check that these identities hold when N is prime, and when $N = LM$ for odd integers L and M in general. □

These properties are useful, but what made the Legendre symbol powerful is quadratic reciprocity. Does this carry over to Jacobi symbols? In fact, the answer is yes!

Theorem 2.16 (Law of Quadratic Reciprocity for Jacobi symbols)

Let $M, N \in \mathbb{N}$ be odd. Then

$$\left(\frac{M}{N}\right) = (-1)^{\left(\frac{M-1}{2}\right)\left(\frac{N-1}{2}\right)} \left(\frac{N}{M}\right)$$

Furthermore, if M and N are coprime, then this means

$$\left(\frac{M}{N}\right) \cdot \left(\frac{N}{M}\right) = (-1)^{\left(\frac{M-1}{2}\right)\left(\frac{N-1}{2}\right)}$$

Proof: Let $M = p_1 \times \cdots \times p_k$ and $N = q_1 \times \cdots \times q_\ell$. Then

$$\left(\frac{M}{N}\right) = \prod_{i=1}^{k}\prod_{j=1}^{\ell}\left(\frac{p_i}{q_j}\right) = \prod_{i=1}^{k}\prod_{j=1}^{\ell}(-1)^{\left(\frac{p_i-1}{2}\right)\left(\frac{q_j-1}{2}\right)}\left(\frac{q_j}{p_i}\right)$$

By combining the products into sums, we see that this is equal to

$$(-1)^{\beta} \times \prod_{i=1}^{k}\prod_{j=1}^{l}\left(\frac{q_j}{p_i}\right) \quad \text{where } \beta = \sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{p_i-1}{2}\right)\left(\frac{q_j-1}{2}\right)$$

where the last term in the multiplication is the Jacobi symbol for N on M.

Now, we must show that the sum is congruent to $\left(\frac{M-1}{2}\right)\left(\frac{N-1}{2}\right)$ modulo 2. But we have previously showed this, so in fact quadratic reciprocity carries over. □

Note: The exponents really are fractions, not Jacobi symbols!

Using these results, we can now compute Jacobi symbols without factorising the numerator:

$$\left(\frac{33}{73}\right) = \left(\frac{73}{33}\right) = \left(\frac{7}{33}\right) = \left(\frac{33}{7}\right) = \left(\frac{5}{7}\right) = \left(\frac{7}{5}\right) = \left(\frac{2}{5}\right) = -1.$$

So the law of quadratic reciprocity, as defined and proved for Legendre symbols in 2.9, is preserved when discussing Jacobi symbols.

3 Quadratic Forms

Our motivating question for this section is "can $n \in \mathbb{N}$ be expressed as $x^2 + y^2$ for integers x, y?" Then, we ask the same question for $x^2 + 2y^2$, $x^2 + 3y^2$, and so on.

Theorem 3.1 (Fermat-Euler Theorem)

Let p be an odd prime. Then the following are equivalent:

1. $p = x^2 + y^2$ for some $x, y \in \mathbb{Z}$.
2. $-1 + p\mathbb{Z} \in (\mathbb{Z}/p\mathbb{Z})^\times$ is a square.
3. $p \equiv 1 \pmod 4$.

More generally, if $N \in \mathbb{N}$, then $N = x^2 + y^2$ if and only if for every p congruent to 3 mod 4, if $p^k \mid N$ but $p^{k+1} \nmid N$, then k is even.

Definition 3.2 (Binary Quadratic Form)

A *binary quadratic form* is a polynomial $f(x, y) = ax^2 + bxy + cy^2$, with a, b, c integers. We then say that f *represents* N if there are integers m, n such that $f(m, n) = N$.

We often identify f with the tuple of coefficients (a, b, c) or the matrix

$$f \sim \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \implies f(x, y) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

We will study these forms and how they behave under a change of variables. We need to restrict which changes are allowed.

Definition 3.3 (Unimodular Change of Variables)

A *unimodular change of variables* is of the form $X = \alpha x + \gamma y$, $Y = \beta x + \delta y$, where

$$\alpha, \beta, \gamma, \delta \in \mathbb{Z} \qquad \alpha\delta - \beta\gamma = 1$$

Equivalently, this is the form $(X, Y) = (x, y)A$, where $A \in \mathrm{SL}_2(\mathbb{Z})$:

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathbb{M}_2(\mathbb{Z}) : \alpha\delta - \beta\gamma = 1 \right\}$$

Two binary quadratic forms f and g are called *equivalent* if there exists a unimodular change of variables such that

$$g(x, y) = f(X, Y) = f(\alpha x + \gamma y, \beta x + \delta y)$$

Remember that the special linear group $\mathrm{SL}_2(\mathbb{Z})$ is indeed a *group*: it is closed under multiplication, matrix multiplication is associative, and inverses are given by

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^{-1} = \begin{pmatrix} \delta & -\beta \\ -\gamma & \alpha \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

This group acts on the set of binary quadratic forms by the formula

$$(Af)(x, y) = f((x, y)A)$$

Then two forms are equivalent if they are in the same orbit under this action. In particular, this demonstrates that equivalence of forms really is an equivalence relation, as orbits partition a set.

3.1 Positive Definite Binary Quadratic Forms

We now consider a property of binary quadratic forms, and relate it to equivalence.

Definition 3.4 (Discriminant)

The *discriminant* of a binary quadratic form $f(x, y) = ax^2 + bxy + cy^2$ is

$$\operatorname{disc} f = b^2 - 4ac$$

Proposition 3.5 (Equivalence of Forms)

Let f and g be equivalent binary quadratic forms. Then

1. f and g represent the same integers.
2. $\operatorname{disc} f = \operatorname{disc} g$.

Proof: If g represents N, then $N = g(m, n) = f(\alpha m + \gamma n, \beta m + \delta n)$, so f represents g too. The converse also holds, since these matrices are invertible in $\mathrm{SL}_2(\mathbb{Z})$.

Now let M_f be the matrix associated with f, so that $\det M_f = ac - b^2/4 = -(\operatorname{disc} f)/4$. Then

$$f(x, y) = (x, y)M_f(x, y)^\top = (x, y)AM_fA^\top(x, y)^\top \implies M_g = AM_fA^\top$$

Thus $\det M_g = \det M_f$, and so $\operatorname{disc} g = \operatorname{disc} f$. □

Note: The discriminants of two forms being the same is not a sufficient condition for them to be equivalent. For example,

$$f(x, y) = x^2 + 6y^2 \qquad g(x, y) = 2x^2 + 3y^2$$

both have discriminant -24, as $\operatorname{disc} f = -4 \cdot 1 \cdot 6 = -4 \cdot 2 \cdot 3 = \operatorname{disc} g$. However f represents 1 via $f(1, 0) = 1$, while g clearly cannot (as the smallest non-zero number it can represent is 2).

Proposition 3.6 (Only Certain Discriminants Possible)

Let $d \in \mathbb{Z}$. Then there exists a binary quadratic form f with $\operatorname{disc} f = d$ if and only if d is congruent to either 0 or 1 modulo 4.

Proof: ($\Rightarrow$) $\operatorname{disc} f = b^2 - 4ac \equiv b^2 \pmod 4$, and the only squares modulo 4 are 0 and 1.

($\Leftarrow$) For d congruent to 0 mod 4, take $f(x, y) = x^2 - (d/4)y^2$.

For d congruent to 1 mod 4, take $f(x, y) = x^2 + xy - ((d - 1)/4)y^2$. □

Definition 3.7 (Definite)

Let $f(x, y)$ be a binary quadratic form. Then

1. f is *positive definite* if for all $(u, v) \in \mathbb{R}^2 \setminus \{0\}$, $f(u, v) > 0$.
2. f is *negative definite* if for all $(u, v) \in \mathbb{R}^2 \setminus \{0\}$, $f(u, v) < 0$.
3. f is *indefinite* if it is non-zero and neither positive nor negative definite.

In particular, every non-zero binary quadratic form is either positive definite, negative definite, or indefinite. From now on, we mostly focus our attention on *positive definite binary quadratic forms*, or PDBQFs for short.

Proposition 3.8 (Definite)

If $f(x,y) = ax^2 + bxy + cy^2$ is a binary quadratic form, with $\operatorname{disc} f = d$. Then

(a) If $d < 0$, then $a \neq 0$. f is positive definite if $a > 0$, and negative definite otherwise.

(b) If $d > 0$, then f is indefinite.

(c) If $d = 0$, then there are integers $l, m, n \in \mathbb{Z}$ such that $f = l(mx + ny)^2$.

Proof: First, notice that

$$\begin{aligned} 4a \cdot f(x,y) &= 4a^2x^2 + 4abxy + 4acy^2 \\ &= (2ax + by)^2 + (4ac - b^2)y^2 = (2ax + by)^2 - dy^2 \end{aligned}$$

(a) If $a = 0$, then $d = b^2 - 4ac = b^2 \not< 0$. Then the RHS of the above expression is positive definite.

(b) If $d > 0$, then the RHS is indefinite, so $f(x,y)$ is too. The same holds for $c \neq 0$ and $d > 0$.

In the case where $a = c = 0$, f is clearly indefinite, as $b \neq 0$.

(c) If $d = 0$, then $b^2 = 4ac$. Write $a = a_1(a_2)^2$, where a_1 is squarefree.

Then $b^2 = 4a_1a_2^2c$, so we have $(2a_2)^2 \mid b^2$ and thus $2a_2 \mid b$.

But then $(b/2a_2)^2 = a_1c$, and so $a_1 \mid (b/2a_2)^2$ (as a_1 is squarefree). Then $2a_1a_2 \mid b$, so

$$f(x,y) = a_1a_2^2x^2 + bxy + cy^2 = \underbrace{a_1\left(a_2x + \frac{b}{2a_1a_2}y\right)^2}_{\text{desired form}} + \underbrace{\left(c - \frac{b^2}{4a}\right)}_{=0}y^2$$

In the case where $a = 0$, then $b = 0$, so $f(x,y) = cy^2 = 1(0x + cy)^2$. □

Now, we turn our attention to PDBQFs. If $f(x,y) = ax^2 + bxy + cy^2$ is a PDBQF, can we find an equivalent form with smaller coefficients? More generally, is there some canonical representative for the equivalence class of f?

Example 3.9 (Reducing Coefficients)

Take $f(x,y) = 10x^2 + 34xy + 29y^2$, or $(1, 34, 29)$. Consider the actions of various elements of $\mathrm{SL}_2(\mathbb{Z})$ on f.

If $T_\lambda = \begin{pmatrix} 1 & 0 \\ \lambda & 0 \end{pmatrix}$, then $(T_\lambda \cdot f)(x,y) = ax^2 + (b + 2\lambda a)xy + (c + \lambda b + \lambda^2 a)y^2$. So

$$\begin{aligned} T_{+1} &: (a,b,c) \mapsto (a, b + 2a, c + b + a) \\ T_{-1} &: (a,b,c) \mapsto (a, b - 2a, c - b + a) \end{aligned}$$

Using these matrices repeatedly, we can get

$$(10, 34, 29) \xrightarrow{T_{-1}} (10, 14, 5) \xrightarrow{T_{-1}} (10, -6, 1)$$

Also, we can consider $S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$. Notice that $S : (a,b,c) \mapsto (c, -b, a)$. Now

$$(10, -6, 1) \xrightarrow{S} (1, 6, 10) \xrightarrow{T_{-1}} (1, 4, 5) \xrightarrow{T_{-1}} (1, 2, 2) \xrightarrow{T_{-1}} (1, 0, 1)$$

Thus the PDBQF $f(x,y) = 10x^2 + 34xy + 29y^2$ is equivalent to $x^2 + y^2$.

Note: We say that f is *reduced* if $-a < b \leqslant a \leqslant c$, and if $a = c$ then $b \geqslant 0$. Equivalently, f is reduced if $|b| \leqslant a \leqslant c$, and if either inequality is an equality, then $b \geqslant 0$.

Proposition 3.10 (Reduction is Possible)

Every PDBQF is equivalent to a reduced PDBQF.

Proof: Start with $f = (a, b, c)$ and consider the following algorithm:

1. If $a > c$, replace f by $S \cdot f$.
2. If $|b| > a$, replace f by $T_{\pm 1} \cdot f$, depending on the sign of b.
3. Keep doing this until you satisfy the conditions.

This terminates in a finite number of steps, as $a + |b|$ decreases with each step.

After running this algorithm, notice that $a \leqslant c$ as required, and $|b| \leqslant a \leqslant c$. We are then done, except in the cases where $|b| = a$ or $a = c$. If $a = c$, then $f = (a, b, a)$, so if $b < 0$ we can take $S \cdot f = (a, -b, a)$. Otherwise, either $f = (a, a, c)$ is reduced or $f = (a, -a, c)$ is not reduced. In the latter case, take $T_1 \cdot f$, which is reduced. □

Proposition 3.11 (Reduced PDBQF Inequalities)

If $f = (a, b, c)$ is a reduced PDBQF, then

$$|b| \leqslant a \leqslant \sqrt{\frac{1}{3}\,|\operatorname{disc} f|} \qquad \text{and} \qquad b \equiv \operatorname{disc} f \pmod 2$$

Proof: Since f is reduced, $|b| \leqslant a \leqslant c$. Thus

$$|\operatorname{disc} f| = 4ac - b^2 \geqslant 4a^2 - a^2 = 3a^2 \implies a \leqslant \sqrt{\frac{1}{3}\,|\operatorname{disc} f|}$$

Also, $\operatorname{disc} f \equiv b^2 \equiv b$ modulo 2. □

Suppose $f = (a, b, c)$ is a reduced PDBQF of discriminant -4. Then $|b| \leqslant a \leqslant \sqrt{4/3}$, so $a = 1$ and b is even, so $b = 0$. Then $d = -4 = b^2 - 4ac = -4c$, so $c = 1$, and $f(x, y) = x^2 + y^2$. This means that there is only one reduced PDBQF of discriminant -4!

Corollary: Since every PDBQF is equivalent to some reduced form, and equivalent forms have the same discriminant, every PDBQF of discriminant -4 is equivalent.

Proposition 3.12 (Represented Primes)

If p is a prime congruent to 1 modulo 4, then p is represented by $x^2 + y^2$.

Proof: Since $p \equiv 1$ modulo 4, we have

$$\left(\frac{-1}{p}\right) = 1 \implies \exists\, k, l \in \mathbb{Z} \text{ s.t. } k^2 = -1 + lp$$

This means that $(2k)^2 = -4 + 4lp$, so $-4 = (2k)^2 - 4lp$.

The PDBQF $f = (p, 2k, l) = px^2 + 2kxy + ly^2$ has $\operatorname{disc} f = -4$. It then has an equivalent reduced form of discriminant -4, which must be $x^2 + y^2$ by the above corollary.

But $f(1, 0) = p$, so f represents p. So $x^2 + y^2$ does too, as desired. □

This is a surprising result! Every prime number which is one more than a multiple of 4 is the sum of two squares.

Corollary: If $d < 0 \in \mathbb{Z}$ is congruent to 0 or 1 modulo 4, there are only finitely many reduced PDBQFs of discriminant d.

Proof: For f reduced, there are only finitely many possible a, b. But then c is fixed. □

Corollary: If $d < 0 \in \mathbb{Z}$ is congruent to 0 or 1 modulo 4, there are only finitely many equivalence classes of PDBQFs of discriminant d.

Proof: Obvious by the previous corollary and every class having a reduced representative. □

Definition 3.13 (Class Number)

For negative integers $d \equiv 0$ or 1 mod 4, we define the *class number* $h(d)$ to be the number of equivalence classes of PDBQFs with discriminant d.

We have computed this to be 1 in the case of $d = -4$. In fact, $h(d) \geqslant 1$ for all d: choose $x^2 - \frac{d}{4}y^2$ or $x^2 + xy + \frac{1-d}{4}y^2$ as appropriate.

Definition 3.14 (Proper Representation)

An integer $N \in \mathbb{Z}$ is *properly represented* by a binary quadratic form f if there are $m, n \in \mathbb{Z}$ with $(m, n) = 1$ such that $f(m, n) = N$.

Note: This is the same as the original definition of representation given in 3.2, with the added stipulation that the integers are coprime.

In fact, the properties of representation carry over nicely to proper representation!

Proposition 3.15 (Equivalence of Forms 2)

As well as the properties given in Proposition 3.5, equivalent binary quadratic forms properly represent the same integers too.

Proof: Suppose we can write $g = A \cdot f$, where $A \in \mathrm{SL}_2(\mathbb{Z})$, and that $g(m, n) = N$. We want to show that f properly represents N as well, which will complete the proof by symmetry.

We get $N = g(m, n) = (A \cdot f)(m, n) = f(\alpha m + \gamma n, \beta m + \delta n)$. We need to check that this is indeed a proper representation: that is, $(\alpha m + \gamma n, \beta m + \delta n) = 1$.

This is, by definition, $(m, n)A$. But then $(m, n) = (\alpha m + \gamma n, \beta m + \delta n)A^{-1}$, since $\mathrm{SL}_2(\mathbb{Z})$ is a group and thus has inverses. Thus if any d divides both $\alpha m + \gamma n$ and $\beta m + \delta n$, then it must also divide both m and n, since these are linear combinations of $\alpha m + \gamma n$ and $\beta m + \delta n$.

But m and n are coprime, so $\alpha m + \gamma n$ and $\beta m + \delta n$ are too. Thus f properly represents N, so proper representation is an equivalence class property. □

Now, we prove some more properties of the values which reduced PDBQFs take.

Proposition 3.16 (Proper Reduction)

Let $f = (a, b, c)$ be a reduced PDBQF. Then:

(i) $a \leqslant c \leqslant a + c - |b|$.

(ii) $f(1, 0) = a$ and $f(0, 1) = c$.

(iii) Either $f(1, 1) = a + c - |b|$ or $f(1, -1) = a + c - |b|$.

(iv) If m and n are non-zero integers, then $f(m, n) \geqslant a + c - |b|$.

Proof: (i) As f is reduced, we have $c \geqslant a \geqslant |b|$. This means $a - |b| \geqslant 0$, and so $a + c - |b| \geqslant c$.

(ii) $f(x, y) = ax^2 + bxy + cy^2$, so $f(1, 0) = a + 0 + 0$ and $f(0, 1) = 0 + 0 + c$ as required.

(iii) Also, $f(1, \pm 1) = a \pm b + c$. For one of these values, we therefore obtain $a + c - |b|$.

(iv) Suppose first that $|m| \geqslant |n|$. Then $f(m, n) = am^2 + bmn + cn^2$. This is at least as large as $am^2 - |b|\, m^2 + cn^2$, which is equal to $(a - |b|)m^2 + cn^2$.

But since m and n are non-zero integers, their squares are at least 1. Therefore $f(m, n) \geqslant a+c-|b|$ whenever $|m| \geqslant |n|$. A similar argument works in the opposite case. □

Note: The number of pairs m and n with $g(m, n) = N$ is finite. It is also even, by symmetry: we can consider $g(-m, -n)$. We can thus take the values taken by g with multiplicity in some order, or indeed all the values taken when m and n are coprime. This is a non-decreasing sequence where each integer appears an even number of times. The interpretation of the above is then that if g is reduced, this list will be a, a, c, c, $a + c - |b|$, $a + c - |b|$, and then more.

Now, recall Proposition 3.10, which stated that every PDBQF was equivalent to a reduced form. In fact, we can strengthen this claim: the reduced form is unique!

Theorem 3.17 (Unique Reduction Theorem)

Every PDBQF is equivalent to a unique reduced form.

Proof: From Proposition 3.10, every PDBQF is equivalent to a reduced form. As equivalence of forms is an equivalence relation, if this form was not unique, then there would exist two distinct reduced forms which were equivalent to each other. We therefore need to show that if $f = (a, b, c)$ and $g = (a', b', c')$ are equivalent reduced forms, then they must be equal.

Take the lists of properly represented integers, as described in the above note. These will be the same for f and g. But then these lists begin the same way, so $a = a'$, and $c = c'$, and (since they must have the same discriminant), $b = \pm b'$.

If $b = 0$, then we are done, so suppose $b \neq 0$. In particular, suppose without loss of generality that $b' < 0 < b$. Then $f = (a, b, c)$ and $g = (a, -b, c)$ are equivalent reduced forms. In particular, the inequalities $c > a > b$ are strict.

So $g(1, 0) = a = (A \cdot f)(1, 0) = f((1, 0)A)$. By the previous proposition, the only way for this to be true is if $(1, 0)A = (\pm 1, 0)$. The same logic shows that $(0, 1)A = (0, \pm 1)$. So:

$$A = \begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix} \text{ where } \det A = 1 \implies A = \pm I \text{ the identity matrix.}$$

But $\pm I$ does not change a form, since it is quadratic! But then $g = f$, as required. □

Corollary: The class number (Definition 3.13) $h(d)$ is the number of reduced PDBQFs with a discriminant of d, yielding an efficient method to compute $h(d)$.

Example 3.18 (Computing Class Numbers)

Let's find $h(-24)$, noting that $-24 \equiv 0 \pmod 4$, by enumerating the reduced forms (a, b, c) of discriminant d. These forms have b even, and $a \leqslant \sqrt{24/3} < 3$, with $b^2 - 4ac = -24$.

If $a = 1$, then we must have $b = 0$. Then $-4c = -24$, so $c = 6$, yielding $(1, 0, 6)$.

If $a = 2$, then $b = 0$ or 2: $b = -2$ is not allowed, as this would not be reduced. For $b = 0$, we have $-8c = 24$, so $c = 3$. For $b = 2$, we have $-8c = -28$, which has no integer solution.

Thus the only reduced forms of discriminant -24 are $(1, 0, 6)$ and $(2, 0, 3)$: that is, $x^2 + 6y^2$ and $2x^2 + 3x^2$. In particular, the class number $h(-24) = 2$.

Proposition 3.19 (Proper Representation Condition)

Let f be a binary quadratic form, and $N \in \mathbb{Z}$. Then f properly represents N if and only if f is equivalent to a form (N, b, c) for some integers b and c.

Proof: ($\Leftarrow$) Proper representation is preserved under equivalence. But $Nx^2 + bxy + cy^2$ properly represents N, for example with $(1, 0)$.

($\Rightarrow$) Suppose $f(m, n) = N$ for coprime m and n. Then we can find u and v with $um + vn = 1$. So:

$$A = \begin{pmatrix} m & n \\ -v & u \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \quad (\text{since } \det A = um + vn = 1).$$

But if $g = A \cdot f$, then $g(1, 0) = f((1, 0)A) = f(m, n) = N$ by assumption. □

Theorem 3.20 (Proper Representation)

Let $d \in \mathbb{Z}$ with $d < 0$ and $d \equiv 0$ or 1 modulo 4 be some valid discriminant. Then the following conditions are equivalent:

1. N is properly represented by a PDBQF of discriminant d.
2. The equation $X^2 = d$ has a solution in $\mathbb{Z}/4N\mathbb{Z}$.

Proof: ($1 \Rightarrow 2$) By the above proposition, there exists some form (N, b, c) of discriminant d. But then $b^2 - 4Nc = d$ by definition, so $b^2 \equiv d$ modulo $4N$ as required.

($2 \Rightarrow 1$) Suppose we have b with $b^2 \equiv d$ modulo $4N$, so $b^2 = d + 4NC$. Then the form (N, b, c) has discriminant d. Also, this form is positive definite, as $d < 0 < N$. □

Let's use this theorem to study a particular PDBQF.

Example 3.21 ($x^2 + xy + 2y^2$)

Which integers are represented by $x^2 + xy + 2y^2$? This form has discriminant $d = 1 - 8 = -7$, so we must find the reduced forms of discriminant -7.

These have b odd, and $|b| \leqslant a \leqslant \sqrt{7/3} < 2$, so $a = |b| = 1$. Since the form is reduced, $b = 1$, and so $c = 2$: in fact, this is our original form! Therefore this is the *only* reduced form with discriminant -7, and any two forms of this discriminant are equivalent.

So by the above theorem, N is properly represented by $x^2 + xy + 2y^2$ if and only if $X^2 \equiv -7$ has a solution modulo $4N$.

Suppose $N = p$ is prime. If $p = 2$, we want $X^2 \equiv 1 \pmod 8$. This works when $X = 3$, so yes, 2 is represented. (Trivially, $(0, 1)$ represents this value.)

If p is odd, then the Chinese Remainder Theorem (1.14) gives an equivalent condition: we need to solve $X^2 \equiv -7$ modulo 4 and modulo p. This works modulo 4, so we just need some solution to $X^2 \equiv -7 \pmod p$.

Obviously, this works if $p = 7$. For $p \neq 7$, we require the Legendre symbol $(-7/p) = 1$, which by quadratic reciprocity (Theorem 2.9) is $(p/7)$, to be 1.

Equivalently, we require $p = 0$, 1, 2, or 4 modulo 7.

Note: In this example, we answered the question for $N = p$ a prime. We now build up to being able to do this for arbitrary integers, which are not necessarily prime.

Proposition 3.22 (Proper Representation and Legendre Symbols)

Let p be an odd prime with $a \in \mathbb{Z}$. Then if the Legendre symbol $(a/p) = 1$, then $X^2 = a$ has a solution in $\mathbb{Z}/p^k\mathbb{Z}$ for all k.

Moreover, if $a \equiv 1 \pmod 8$, then the equation $X^2 = a$ has a solution in $\mathbb{Z}/2^k\mathbb{Z}$ for all k.

Proof: We use induction on $k \geqslant 1$, since the base case of $k = 1$ holds by definition of the Legendre symbol. Suppose that there is some $b \in \mathbb{Z}$ with $b^2 \equiv a \pmod{p^k}$, or equivalently that there is some $c \in \mathbb{Z}$ with $b^2 = a + p^k c$.

Then $(b+p^k x)^2 = b^2 + 2bp^k x + p^{2k}x^2 = a + p^k(c+2bx) + p^{2k}x^2$. This last term is clearly a multiple of p^{k+1}, so in fact we merely require $p \mid (c + 2bx)$. Equivalently, if $2bx \equiv -c \pmod p$. This is possible if $(p, 2b) = 1$. But this is true, since $p \nmid 2$ and $p \nmid b$, since $a \not\equiv 0 \pmod p$ by the Legendre symbol, and so $b \not\equiv 0 \pmod p$. This proves the first part of the proposition.

Now we use induction again. For $k \leqslant 3$, this has a solution by assumption. Suppose $b \in \mathbb{Z}$ is such that $b^2 = a \pmod{2^k}$, where k is at least 3. Then there exists some $c \in \mathbb{Z}$ with $b^2 = a + 2^k c$. If c is even, then $b^2 \equiv a \pmod{2^{k+1}}$ as required, so take c odd.

Then b is odd, since $b^2 \equiv a \equiv 1 \pmod 2$, so $(b+2^{k-1})^2 = b^2 + 2^k b + 2^{2k-2} = a + 2^k(b+c) + 2^{2k-2}$. But b and c are odd, so $2^k(b+c) \equiv 0 \pmod{2^{k+1}}$. Then we are done, provided that $2k-2 \geqslant k+1$, which is indeed true for $k \geqslant 3$. □

Corollary: $N \in \mathbb{N}$ is properly represented by $x^2 + xy + 2y^2$ if and only if for every prime $p \mid N$, we do not have $p \equiv 3$, 5, or 6 modulo 7, and we do not have $49 \mid N$.

We are often interested in representation in general, rather than specifically proper representation. Suppose that $f(m, n) = N$, with $(m, n) = k$. Then $m = km'$ and $n = kn'$, and $f(m, n) = k^2N'$, where $N' = f(m', n')$. In particular, N' is properly represented by f.

Corollary: $N \in \mathbb{N}$ is represented by $x^2 + xy + 2y^2$ if and only if every prime $p \mid N$ with $p \equiv 3$, 5, or 6 modulo 7 is such that the highest power p^k dividing N has k even.

Remark 3.23 (Difficulty of Characterisation)

We know that if $d < 0$ is congruent to 0 or 1 modulo 4, and $h(d) = 1$, we can characterise which natural numbers $N \in \mathbb{N}$ are represented by the unique reduced PDBQF of discriminant d in terms of congruence conditions on the primes $p \mid N$. This generalises the Fermat-Euler Theorem (3.1).

If $h(d) > 1$, then this method isn't quite as precise. We can only characterise the integers $N \in \mathbb{N}$ which are represented by *some* PDBQF of discriminant d, but we can't always easily tell which. Are we simply missing something? In fact, no. One can show that congruence conditions generally do not suffice to characterise the primes p represented by a given PDBQF!

The form $f(x, y) = x^2 + xy + 6y^2$ has discriminant $d = -23$, and $h(-23) = 3$, so we can't find the primes represented by f easily, but we *can* show that any prime $q \neq 23$ is represented by f if and only if the coefficient of r^q in the product

$$r \prod_{n=1}^{\infty} \left((1 - r^n) \times (1 - r^{23n})\right)$$

is equal to 2. This is a strange result, which goes far beyond the scope of this course.

The class numbers $h(d)$ have been well-studied. For example, Siegel and Heilbronn proved in 1934 that as $d \to -\infty$, $h(d) \to \infty$. Additionally, Baker and Stark proved in 1967 that the only discriminants d with a unique reduced form (that is, with $h(d) = 1$) are:

$$-3,\ -4,\ -7,\ -8,\ -11,\ -19,\ -43,\ -67,\ \text{and}\ -163.$$

4 The Distribution of Primes

At the very beginning of this course, in §**??**, we considered questions about the prime numbers. In this section, we are interested in questions along the lines of "what is the probability that a randomly selected 50-digit integer is prime?". This is highly useful in cryptography: for example, we may want to efficiently generate RSA numbers $N = pq$.

One method we have considered to find large prime numbers is to test random numbers, and see if they are prime. This method will be more efficient the higher the density of primes in this range is. This density is obviously given by the prime counting function:

$$\text{prime density} = \frac{\pi(10^{50}) - \pi(10^{49})}{10^{50} - 10^{49}} \text{ where } \pi(x) = \#\left\{p \mid 1 \leqslant p \leqslant x, p \text{ a prime}\right\}.$$

So we want to study the behaviour of the prime counting function $\pi(x)$.

Theorem 4.1 (Prime Number Theorem)

$\pi(x) \sim x/\log x$, where $\log = \log_e = \ln$ is the natural logarithm.

Note: If f and $g : (0, \infty) \to (0, \infty)$, say $f \sim g$ if the limit of $f(x)/g(x)$ is 1 as $x \to \infty$.

One can show that $x/\log x \sim \text{li}(x)$, where li is the *logarithmic integral* as given in §**??**:

$$\text{li}(x) = \int_2^x \frac{1}{\log t}\, dt.$$

In fact, li is a better approximation to $\pi(x)$ than $x/\log x$ for large x. This means that the density of the primes around x is, in the limit, around $1/\log x$.

Corollary: The probability that a random 20-digit integer is prime is around $1/\log(5 \times 10^{19})$, which is around 0.02205, or just under one in 45. In fact, the true value, known by testing every 20-digit number, is around 0.0220, so the approximation is very accurate!

There are many different formulations of the Prime Number Theorem.

Theorem 4.2 (Dirichlet's Theorem on Primes in Arithmetic Progressions)

If $a \in \mathbb{Z}$ and $N \in \mathbb{N}$ with $(a, N) = 1$, there are infinitely many primes $p \in a + N\mathbb{Z}$.

This is equivalent to the Prime Number Theorem by an alternative statement, which says that:

$$\pi(a, N, x) = \#\left\{\text{primes } p \leqslant x : p \equiv a \pmod N\right\} \sim \frac{1}{\phi(x)} \times \frac{x}{\log x}$$

with ϕ being Euler's totient function, and in particular that the limit

$$\lim_{x \to \infty} \frac{\pi(a, N, x)}{\pi(x)} = \frac{1}{\phi(N)}.$$

So the primes are uniformly distributed among the possible classes in $(\mathbb{Z}/N\mathbb{Z})^\times$.

Unfortunately, we will not be able to prove the Prime Number Theorem or Dirichlet's Theorem, as they require a lot of technical work beyond the scope of this course. However, we will discuss the Riemann zeta function $\zeta(s)$, which is used in the proof of the Prime Number Theorem, and we will give a proof of Chebyshev's theorem, which states that there are $0 < c_1 \leqslant c_2$ with

$$c_1 x/\log x \leqslant \pi(x) \leqslant c_2 x/\log x \quad \text{for all } x \geqslant 2.$$

Proposition 4.3 (First Prime Counting Bound)

If $x \in \mathbb{Z}$ with $x \leqslant 2$, then $\pi(x) \geqslant \log x / 2 \log 2$.

Proof: We are going to think of x as being the number of integers between 1 and x, so that we have $[x] = \#\{1, 2, \ldots, x\}$, and consider an alternate way of counting this set. Let $p_1, \ldots, p_r$ be the primes $\leqslant x$, so that $\pi(x) = r$. Any $1 \leqslant N \leqslant x$ can be written uniquely in the form:

$$N = p_1^{a_1} \times p_1^{a_2} \times \cdots \times p_r^{a_r} \times M^2$$

where $a_i \in \{0, 1\}$, and $M^2 \leqslant N \leqslant x$, so in particular $M \leqslant \sqrt{x}$. But now we can count the elements in $[x]$ quite easily: there are two choices for each p_i, and at most $\sqrt{x}$ choices for M. Thus:

$$x \leqslant 2^r \sqrt{x} \implies 2^{\pi(x)} \geqslant \sqrt{x}$$

Taking logarithms on both sides yields $\pi(x) \geqslant \log \sqrt{x} / \log 2$, proving the result. □

Proposition 4.4 (Prime Sum Diverges)

The infinite sum and infinite product over all primes:

$$\sum_{p \text{ a prime}} 1/p \qquad \text{and} \qquad \prod_{p \text{ a prime}} \left(1 - \frac{1}{p}\right)^{-1}$$

both diverge. That is, the sequence of partial sums and products diverge.

Proof: We first show that the two divergences are equivalent by using the Taylor series expansion:

$$-\log(1 - x) = \sum_{k=1}^{\infty} \frac{x^k}{k} \quad \text{(which is absolutely convergent for } |x| < 1)$$

We then take the logarithm of the (finite truncation of) the product, which is:

$$\log \prod_{p \leqslant x \text{ a prime}} \left(1 - \frac{1}{p}\right)^{-1} = \sum_{p \leqslant x \text{ a prime}} -\log(1 - 1/p) = \sum_{p \leqslant x \text{ a prime}, k \geqslant 1} p^{-k}/k$$

This sum, if we take $k \geqslant 2$ instead of $k \geqslant 1$, is bounded as $x \to \infty$. But the sum can in fact be split into the sum with $k \geqslant 2$ and the sum with $k = 1$, and the sum when $k = 1$ is specifically the sum of $p^{-1}/1 = 1/p$ over all primes p. This is the sum we want to show diverges!

First, we prove that the sum for $k \geqslant 2$ really is bounded, using:

$$\sum_{p \leqslant x \text{ a prime}, k \geqslant 2} p^{-k}/k \leqslant \sum_{p \leqslant x \text{ a prime}, k \geqslant 2} p^{-k} \leqslant \sum_{p \leqslant x \text{ a prime}} \frac{1}{p^2} \times \frac{1}{1 - 1/p} \leqslant \sum_{p \leqslant x \text{ a prime}} \frac{1}{p(p-1)}$$

But this sum is bounded by the same sum but with any number $n \geqslant 2$ rather than specifically a prime p, and this is bounded by the sum of $1/n^2$ plus some constant for $n = 2$. This is bounded by $\pi^2/6$, and so the original sum is finite, as we require.

Now, we prove that the infinite product diverges. Let $p_1, \ldots, p_r$ be the primes less than or equal to x, so that we consider the truncated product. Then we see that:

$$\prod_{p \leqslant x \text{ a prime}} (1 - 1/p)^{-1} = \prod_{p \leqslant x \text{ a prime}} \sum_{k \in \mathbb{N}} p^{-k} = \sum_{k_1, \ldots, k_r \geqslant 0} (p_1^{k_1} \times p_2^{k_2} \times \cdots \times p_r^{k_r})^{-1}.$$

But any integer $N \leqslant x$ is a product of primes below x, so this is at least the harmonic series! We know that this diverges, so in fact the original product must diverge too. □

4.1 The Riemann Zeta Function

The *Riemann Hypothesis* is perhaps the most famous unsolved problem in mathematics: it is one of the seven Millennium Prize Problems, and so comes with a million-dollar prize from the Clay Mathematics Institute. This problem considers the behaviour of the *Riemann Zeta function*, which is intimately connected to the distribution of primes.

Definition 4.5 (Riemann Zeta Function)

The *Riemann Zeta function* is the function $\zeta : \mathbb{C} \to \mathbb{C}$ given by:

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}.$$

This function was first studied by Euler for $s \in \mathbb{R}$, and then later by Bernhard Riemann, who in 1859 extended the study of ζ to $\mathbb{C}$.

Note: From now on, for $s \in \mathbb{C}$, we typically write $s = \sigma + it$ for $\sigma, t \in \mathbb{R}$.

Proposition 4.6 (Riemann Convergence)

Let $s \in \mathbb{C}$ have real part $\mathrm{Re}(s) > 1$. Then the series defining $\zeta(s)$ converges absolutely.

Proof: We can simply evaluate the sum for $s = \sigma + it$ with $\sigma > 1$:

$$\sum_{n=1}^{\infty} \left|n^{-s}\right| = \sum_{n=1}^{\infty} |\exp(-\sigma \log n) \cdot \underbrace{\exp(-it \log n)}_{\text{magnitude 1}}| = \sum_{n=1}^{\infty} \exp(-\sigma \log n) = \sum_{n=1}^{\infty} n^{-\sigma}.$$

Here, all the summands are positive real numbers, and we know that this series converges absolutely if and only if $\sigma > 1$. Thus the original series for $\zeta(s)$ does too. □

Corollary: In fact, we have proved something even stronger! The same argument shows that $\zeta(s)$ converges *uniformly* within any half-plane in $\mathbb{C}$ of the form $\{\mathrm{Re}(z) \geqslant 1 + \delta : z \in \mathbb{C}\}$, where $\delta > 0$. Moreover, the uniform limit of holomorphic functions is itself holomorphic, so in fact $\zeta(s)$ must be holomorphic on $\{s \in \mathbb{C} : \mathrm{Re}(s) > 1\}$.

Proposition 4.7 (Euler Product)

Let $s \in \mathbb{C}$ have real part $\sigma > 1$. Then:

$$\zeta(s) = \prod_{p \text{ prime}} \left(1 - p^{-s}\right)^{-1}.$$

Moreover, this product is not equal to 0.

Proof: Formally, we can expand this infinite product as an infinite series:

$$\prod_{p \text{ prime}} \left(1 - p^{-s}\right)^{-1} = \prod_{p \text{ prime}} \left(1 + p^{-s} + p^{-2s} + p^{-3s} + \dots\right)$$

We can regroup the terms in this product to be over selections of r distinct primes:

$$\sum_{r \geqslant 0} \sum_{p_1 < \cdots < p_r \text{ primes}} \sum_{k_1, \dots, k_r \geqslant 1} \left(p_1^{k_1} \times \cdots \times p_r^{k_r}\right) = \sum_{n=1}^{\infty} n^{-s},$$

where the last inequality follows by the fact that each integer can be represented exactly once as the product of distinct primes raised to non-zero powers.

Unfortunately, this result does not itself immediately prove the proposition, as the product might not necessarily converge. We must argue slightly more rigorously. Take $X > 2$ and consider the primes $p_1, \ldots, p_r$ which are at most X: indeed, $r = \pi(X)$. Then:

$$\prod_{i=1}^{r} \left(1 - p_i^{-s}\right)^{-1} = \prod_{i=1}^{r} \left(1 + p_i^{-s} + p_i^{-2s} + \ldots\right) = \sum_{n \in S_X} n^{-s},$$

where S_X is the set of numbers whose prime factors are all at most X. But the difference between this partial product and $\zeta(s)$ is at most the sum over the numbers not in S_X:

$$D_X = \left|\zeta(s) - \prod_{p \leqslant X} (1 - p^{-s})^{-1}\right| = \left|\zeta(s) - \sum_{n \in S_X} n^{-s}\right| \leqslant \sum_{n \in \mathbb{N} \setminus S_X} \left|n^{-s}\right| \leqslant \sum_{n > X} n^{-\sigma}.$$

But the series on the right must converge to 0 as $X \to \infty$, since the full sum converges for $\sigma > 1$. Therefore the error term D_X does too, and so in fact the Euler product is equal to $\zeta(s)$.

To show that this product does not vanish, we consider:

$$\left|\zeta(s) \prod_{p \leqslant X} \left(1 - p^{-s}\right)\right| = \left|\prod_{p > X} \left(1 - p^{-s}\right)^{-1}\right| = \left|1 + \sum_{n \in T_X} n^{-s}\right|,$$

where T_X is the set of numbers n such that all prime factors of n are greater than X. In particular, $T_X \subseteq \{n > X : n \in \mathbb{N}\}$, so this sum is bounded by the sum with $n > X$.

For sufficiently large X, as the tail sum converges to 0, we must have:

$$\left|\zeta(s) \prod_{p \leqslant X} \left(1 - p^{-s}\right)\right| \geqslant 1 - \sum_{n > X} n^{-s} > 0$$

which of course can only happen if $\zeta(s) \neq 0$, as desired. □

Remark 4.8 (Properties of ζ)

We proved earlier that $\zeta(s)$ is holomorphic and does not vanish in the half-plane defined by $\{\mathrm{Re}(s) > 1 : s \in \mathbb{C}\}$. In fact, it has a meromorphic continuation to $\mathbb{C}$, with a unique simple pole at $s = 1$.

There is also a functional equation relating $\zeta(s)$ and $\zeta(1-s)$. First, recall the *Gamma function*, which for $s \in \mathbb{C}$ with $\mathrm{Re}(s) = \sigma > 0$ is given by

$$\Gamma(s) = \int_0^\infty e^{-t}\, t^{s-1}\, dt.$$

In fact, this function also has a meromorphic continuation to all of $\mathbb{C}$, and we usually consider this continuation to be Γ. This does not vanish, and has simple poles only at the non-positive integers $\{0, -1, -2, \ldots\}$.

We define the *completed* ζ function to be $\xi(s) = \pi^{-s/2} \times \Gamma(s/2) \times \zeta(s)$. As the product of meromorphic functions, this ξ is also mermorphic, with simple poles only at 0 and 1. Most importantly, it satisfies the functional equation $\xi(s) = \xi(1 - s)$, which can be turned into a functional equation for ζ if desired.

Now, ζ and Γ are both non-vanishing when $\sigma > 1$, so ξ is too. Also, the functional equation yields that ξ is non-vanishing for $\sigma < 0$, except when $\Gamma(s/2)$ has a pole, at the negative even integers! At these points, $\zeta(s)$ must be zero: these are called the *trivial zeros*.

Note: For $\sigma > 1$, there are no zeros, and for $\sigma < 0$, there are only the trivial zeros. The strip in the middle $\{0 \leqslant \mathrm{Re}(s) \leqslant 1 : s \in \mathbb{C}\}$ is called the *critical strip*. There is a close relationship between the zeros of ζ in the critical strip, and the behaviour of $\pi(X)$.

Proposition 4.9 (Riemann Hypothesis)

If $s \in \mathbb{C}$ is a non-trivial zero of $\zeta(s)$, then $\mathrm{Re}(s) = 1/2$.

Proof: Obvious. (Just kidding: this is unproven and merely conjectured, with a million dollars and a life of fame on the table for anyone who can prove or disprove it!) □

Note: In fact, almost all mathematicians strongly believe that this conjecture is true. The first ten billion (10^{13}) zeros have been checked, sorted by the magnitude of their imaginary part, and all of them lie on the line $\mathrm{Re}(s) = 1/2$.

4.2 Dirichlet Series

We now introduce a new and useful class of functions, and a way to combine two of them.

Definition 4.10 (Dirichlet Series)

A *Dirichlet series* is a formal power series of the form

$$\sum_{n=1} a_n \cdot n^{-s}$$

where $(a_n)_{n=1}^{\infty}$ is a sequence of complex numbers.

Note: A Dirichlet series which sets $a_n = 1$ for all n clearly yields the Riemann Zeta function.

If we do not restrict the sequence (a_n), then really a Dirichlet series is nothing more than a formal expression. However, if we restrict $|a_n|$ to grow at most as fast as n^{α} for some fixed α, then indeed the corresponding series converges in some half-plane.

Suppose we have two functions f and $g : \mathbb{N} \to \mathbb{C}$. We can think of these functions as being complex sequences (f_n) and (g_n), where $f_n = f(n)$ and $g_n = g(n)$. But then we can write:

$$\left(\sum_{n=1}^{\infty} f_n \cdot n^{-s}\right)\left(\sum_{m=1}^{\infty} g_m \cdot m^{-s}\right) = \sum_{n,m=1}^{\infty} f(n) \cdot g(m) \cdot (nm)^{-s}$$

Now, we count up how many times each natural number r appears on the right: indeed, it appears once for each divisor $d \mid n$. We can therefore write this Dirichlet product as:

$$\sum_{n=1}^{\infty} h(n) \cdot n^{-s} \text{ where } h(n) = \sum_{d|n} f(d) \cdot g(n/d).$$

Definition 4.11 (Dirichlet Convolution)

For functions f and $g : \mathbb{N} \to \mathbb{C}$, we define the *Dirichlet convolution* $f * g$ to be the expression we manipulated above.

We define $\sigma(n) = (\mathrm{id} * \mathbf{1})(n)$, where id is the identity map $n \mapsto n$ and $\mathbf{1}$ is the constant map $n \mapsto 1$. Considering the above expression, this is the sum over divisors $d \mid n$ of d: that is, the sum of the divisors of n.

Proposition 4.12 (Dirichlet Convolution Properties)

For functions f, g, and $h : \mathbb{N} \to \mathbb{C}$, we have:

1. Commutativity: $(f * g) = (g * f)$.
2. Associativity: $(f * g) * h = f * (g * h)$.
3. Preservation of multiplicativity: if f and g are multiplicative, so is $(f * g)$. Recall the definition of multiplicativity of a function from 1.16.

Proof: (1) Firstly, we may evaluate the definition directly:

$$(f * g)(n) = \sum_{d|n} f(d) \cdot g(n/d) = \sum_{ab=n} f(a) \cdot g(b) = \sum_{d|n} f(n/d) \cdot g(d) = (g * f)(n).$$

(2) A similar calculation yields associativity:

$$((f * g) * h)(n) = \sum_{d|n} (f * g)(d) \cdot h(n/d) = \sum_{d|n} \sum_{e|d} f(e) \cdot g(d/e) \cdot h(n/d) = \sum_{abc=n} f(a) \cdot g(b) \cdot h(c).$$

which is entirely symmetric in f, g, and h.

(3) Now, suppose that f and g are multiplicative, and that $(m, n) = 1$. Then:

$$(f * g)(mn) = \sum_{d|mn} f(d) \cdot g(mn/d) = \sum_{d|m} \sum_{e|n} f(de) \cdot g(mn/de).$$

This follows by each factor $d \mid mn$ being uniquely representable as some factor of m multiplied by a factor of n, as $(m, n) = 1$. But then $(d, e) = 1$, and we can use the multiplicativity of f and g:

$$(f * g)(mn) = \sum_{d|m} \sum_{e|n} f(d) \cdot g(m/d) \cdot f(e) \cdot g(n/e) = (f * g)(m) \cdot (f * g)(n).$$

This proves that $(f * g)$ is multiplicative, as desired. □

Definition 4.13 (Möbius Function)

The *Möbius function* $\mu : \mathbb{N} \to \mathbb{C}$ is defined by:

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 \times \cdots \times p_k, \text{ where these are } k \text{ distinct primes} \\ 0 & \text{otherwise, that is if } m^2 \mid n \text{ for some } m. \end{cases}$$

In particular, $\mu(1) = 1$, since 1 is the empty product of 0 primes.

Proposition 4.14 (Multiplicativity)

The Möbius function μ is multiplicative.

Proof: Consider coprime m and n.

If either of $\mu(m)$ and $\mu(n)$ are zero, then one of m and n is not square-free, and so their product is not square-free either. This means $\mu(mn) = 0 = \mu(m) \cdot \mu(n)$ as required.

However, if both m and n are square-free, then $m = p_1 \times \cdots \times p_k$ for some list of k primes, and likewise $n = q_1 \times \cdots \times q_\ell$ for some list of ℓ primes. These lists must be disjoint, as $(m, n) = 1$. But then their product m_n is the product of $k + \ell$ distinct primes. This also satisfies multiplicativity, as $\mu(mn) = (-1)^{k+\ell} = (-1)^k \cdot (-1)^\ell = \mu(m) \cdot \mu(n)$. □

Proposition 4.15 (Convolution Identity)

Consider a function $f : \mathbb{N} \to \mathbb{C}$, and let $\mathbf{1} : \mathbb{N} \to \mathbb{C}$ be the map $n \mapsto 1$. Define a new function $\delta : \mathbb{N} \to \mathbb{C}$ by $\delta(n) = 1$ if $n = 1$ and $\delta(n) = 0$ otherwise.

Then δ is the identity for Dirichlet convolution: $(f * \delta) = f$. Moreover, δ can be broken down further: in fact, $\mu \cdot \mathbf{1} = \delta$.

Proof: We can see the identity property easily by expanding:

$$(f * \delta)(n) = \sum_{d|n} f(d) \cdot \delta(n/d) = f(n).$$

This is because the only non-zero term is when $d = n$, as otherwise $n/d \neq 1$ and so $\delta(n/d) = 0$.

Now, since μ is multiplicative (Proposition 4.14), and multiplicativity is preserved under Dirichlet convolution (Proposition 4.12), and $\mathbf{1}$ is multiplicative (since $1 \cdot 1 = 1$), $(\mu \cdot \mathbf{1})$ is multiplicative.

Moreover, δ is clearly multiplicative. So to show that $(\mu \cdot \mathbf{1}) = \delta$, we need only show equality when n is a prime power p^k, including 1 (for $k = 0$).

It is easy to check that $(\mu * \mathbf{1})(1) = \mu(1) = 1 = \delta(1)$, as the only divisor of 1 is 1. For non-zero powers k, we find that:

$$(\mu * \mathbf{1})(p^k) = \sum_{d|p^k} \mu(d) \cdot \mathbf{1}(p^k/d) = \sum_{i=0}^{k} \mu(p^i) = \mu(1) + \mu(p) + \mu(p^2) + \cdots + \mu(p^k).$$

But clearly $\mu(1) = 1$, $\mu(p) = -1$, and $\mu(p^k) = 0$ for $k > 1$, by the definition of the function. This means that $(\mu * \mathbf{1})(p^k) = 1 - 1 + 0 = 0 = \delta(p^k)$ for prime powers p^k with $k > 0$. □

Theorem 4.16 (Möbius Inversion Formula)

Suppose f and $g : \mathbb{N} \to \mathbb{C}$ is such that for all $n \in \mathbb{N}$, we have:

$$f(n) = \sum_{d|n} g(d).$$

Then in fact we have $g = (\mu * f)$.

Proof: We have $f = (g * \mathbf{1})$. Thus $\mu * f = \mu * g * \mathbf{1} = g * \mu * \mathbf{1} = g * \delta = g$. □

Definition 4.17 (Chebyshev and von Mangoldt Functions)

The *von Mangoldt function* $\Lambda : \mathbb{N} \to \mathbb{C}$ is defined by:

$$\Lambda(n) = \begin{cases} \log(p) & \text{if } n = p^k \text{ for some prime } p \text{ with } k \geqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

The *Chebyshev ψ-function* $\psi : \mathbb{N} \to \mathbb{C}$ is defined by:

$$\psi(X) = \sum_{1 \leqslant n \leqslant X} \Lambda(n).$$

Note: ψ is similar to the prime counting function π, but counts primes with "weight" $\log(p)$ as opposed to weight 1. In fact, one can show easily that $\psi(X) \sim \pi(X) \log(X)$, so it is sufficient to show that $\psi(X) \sim X$ to prove the Prime Number Theorem (4.1).

Theorem 4.18 (Zeta-Lambda Relation)

If $s \in \mathbb{C}$ has real part $\sigma > 1$, then:

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{n=1}^{\infty} \Lambda(n) \cdot n^{-s}.$$

Proof: We have the Euler product (Proposition 4.7), which gives us the relation:

$$\zeta(s) = \prod_{p \text{ prime}} \left(1 - p^{-s}\right)^{-1}.$$

The expression $-\zeta'(s)/\zeta(s)$ is the logarithmic derivative of $\zeta(s)$. We know that $-\log(1-z)$ has a Taylor series in the open unit disk $\{|z| < 1 : z \in \mathbb{C}\}$, given by:

$$-\log(1-z) = \sum_{k=1}^{\infty} \frac{z^k}{k}.$$

It follows that some branch of $\log(\zeta(s))$ is given in the usual half-plane by:

$$\log(\zeta(s)) = -\sum_{p \text{ prime}} \log(1 - p^{-s}) = -\sum_{p \text{ prime}} \sum_{k=1}^{\infty} p^{-ks}/k.$$

Taking the derivative of this yields:

$$-\frac{d}{ds}\log\zeta(s) = -\frac{\zeta'(s)}{\zeta(s)} = \frac{d}{ds}\sum_{p \text{ prime}}\sum_{k=1}^{\infty} -p^{-ks}/k = \sum_{p \text{ prime}}\sum_{k=1}^{\infty} k\log(p)\cdot p^{-ks}/k = \sum_{n=1}^{\infty}\Lambda(n)\cdot n^{-s},$$

exactly as required. The interchange of differentiation and summation is justified by the fact that we have a locally uniformly convergent sum of holomorphic functions. □

This result implies the Prime Number Theorem (4.1). One may consider a contour integral of $-(\zeta'(s)/\zeta(s)) \times (X^s/s)$, which gives a formula for ψ:

$$\psi(X) = X - \frac{\zeta'(0)}{\zeta(0)} - \sum_{\rho \in Z} \frac{X^\rho}{\rho} \qquad \text{where } Z \text{ is the set of zeros of } \zeta(s).$$

4.3 Bertrand's Postulate

Proposition 4.19 (Legendre's Formula)

Suppose $X > 2$, and let P be the product of all primes which are at most $\sqrt{X}$. Then:

$$\pi(X) - \pi(\sqrt{X}) + 1 = \#\left\{1 \leqslant n \leqslant X : (n, P) = 1\right\} = \sum_{d \mid P} \mu(d) \times \lfloor X/d \rfloor.$$

Proof: If $1 \leqslant n \leqslant X$, then either n is prime, or $n = 1$, or there is a non-trivial factorisation $n = ab$ with $a, b > 1$, where either a or b are at most $\sqrt{X}$ (otherwise their product is greater than X).

So if $n \neq 1$ is not prime, then there is some prime p with $p \mid N$ and $p \leqslant \sqrt{X}$. Thus:

$$\begin{aligned}\{1 \leqslant n \leqslant x : (n, P) = 1\} &= \{1 \leqslant n \leqslant X : \text{if } p \leqslant \sqrt{X} \text{ is prime, then } p \nmid n\} \\ &= \{\sqrt{X} < p \leqslant X : p \text{ is a prime}\} \cup \{1\}\end{aligned}$$

and by definition, the right hand side has size $\pi(X) - \pi(\sqrt{X}) + 1$.

The second equality is proved using inclusion-exclusion. Define $A_d = \#\{1 \leqslant n \leqslant X : d \mid N\}$ for $d \mid P$, and write $p_1, \ldots, p_r$ for those primes. Then the set we have studied is:

$$\{1 \leqslant n \leqslant X : (n, P) = 1\} = A_1 \setminus \left(\bigcup_{i=1}^{r} A_{p_i} \right).$$

The size of this set is therefore equal to:

$$\lfloor X \rfloor + \sum_{i=1}^{r} \sum_{j_1 < \cdots < j_i} (-1)^i \cdot \left| A_{p_{j_1}} \cap \cdots \cap A_{p_{j_i}} \right|$$

The size of each of these intersections is the number of multiples of the product of the indices of the sets which are at most X. Writing d for this product, we get $\lfloor X/d \rfloor$. The $(-1)^i$ counts the number of primes in the prime factorisation of d in the same way the Möbius function does.

Writing the double sum as a single sum therefore yields

$$\#\{1 \leqslant n \leqslant X : (n, P) = 1\} = \sum_{d \mid P} \mu(d) \times \lfloor X/d \rfloor,$$

exactly as required. □

Definition 4.20 (p-adic Valuation)

Let $N \in \mathbb{N}$ and let p be a prime. Then the *p-adic valuation* of N, written $v_p(N)$, is equal to the exponent of the largest power of p which divides N. Equivalently, it is the unique integer v where $N = p^v N_0$, where $(N_0, p) = 1$.

This valuation is zero if and only if $p \nmid N$, and positive otherwise.

Corollary: This behaves much like the logarithm, in that $v_p(NM) = v_p(N) \times v_p(M)$.

We now prove some properties of this valuation as it relates to binomial coefficients. This will be useful in the proof of Chebyshev's theorem.

Proposition 4.21 (Valuation Bound for Binomial Coefficients)

Let $n \in \mathbb{N}$ and define $N = (2n)! \div (n!)^2$, which is equal to $2n$ choose n. Then:

1. $2^{2n}/2n \leqslant N < 2^{2n}$.
2. If p is a prime with $n < p \leqslant 2n$, then $v_p(N) = 1$.
3. If p is an odd prime with $2n < 3p \leqslant 3n$, then $v_p(N) = 0$.
4. If p is any prime, then $p^{v_p(N)} \leqslant 2n$.

Proof: (1) We use the fact that $2^{2n} = (1+1)^{2n}$, which can be expanded into a binomial sum of $2n+1$ terms, one of which is N. Therefore:

$$\frac{2^{2n}}{2n} = \frac{1}{2n}\left(2 + \sum_{i=1}^{2n-1} \binom{2n}{i}\right) \leqslant \frac{1}{2n}\left(2 + (2n-1)N\right) \leqslant N < \sum_{i=0}^{2n} \binom{2n}{i} = 2^{2n}.$$

(2) Clearly, p appears once in $2n!$ and zero times in $n!$. Therefore N divides by p, but not p^2.

(3) Similarly, p appears twice in $2n!$ (as $2n < 3p$ but $2p \leqslant 2n$), and once in $n!$, so twice in $(n!)^2$. Thus these appearances cancel out, and so $p \nmid N$, and we have $v_p(N) = 0$ as required.

(4) We show that if $k \geqslant 1$ with $p^k > 2n$, then $v_p(N) < k$. We have $v_p(N) = v_p(2n!) - 2v_p(n!)$. We now use a formula for the p-adic valuation of a factorial, which we show below:

$$v_p(m!) = \sum_{j=1}^{m} v_p(j) = \sum_{j=1}^{m}\sum_{i=1}^{\infty} \mathbf{1}_{\{p^i | j\}} = \sum_{i=1}^{\infty}\sum_{j=1}^{m} \mathbf{1}_{\{p^i | j\}} = \sum_{i=1}^{\infty} \lfloor m/p_i \rfloor .$$

Using this formula for $v_p(N) = v_p(2n!) - 2v_p(n!)$ yields:

$$v_p(N) = \sum_{i=1}^{\infty} \left\lfloor 2n/p^i \right\rfloor - 2\left\lfloor n/p^i \right\rfloor .$$

Since $p^k > 2n$, the terms with $i \geqslant k$ all vanish, while the other terms are at most 1, since for *any* real number x, $\lfloor 2x \rfloor - 2\lfloor x \rfloor \in \{0, 1\}$, because if $x = y + \alpha$ for $\alpha \in [0, 1)$, this expression is equal to $\lfloor 2\alpha \rfloor$ where $2\alpha < 2$. The largest value this sum can take is therefore $k - 1$.

But then this proves our result, as $v_p(N) < k$ as required. □

With this proposition proved, we are ready to prove Chebyshev's theorem! This is a slightly weaker claim than the Prime Number Theorem (4.1) but it is still an interesting and powerful result.

Theorem 4.22 (Chebyshev's Theorem)

There exist positive constants c_1 and c_2 such that for all $X > 4$, we have:

$$c_1 \times \frac{X}{\log X} \leqslant \pi(x) \leqslant c_2 \times \frac{X}{\log X}.$$

In fact, we can take $c_1 = \frac{1}{2}\log(2)$ and $c_2 = 6\log(2)$.

Proof: (Upper bound). We will prove the statement for certain integer values of X, then use the properties of the function $X \mapsto X/\log(X)$ to fill in the gaps. We begin our proof by claiming that for $k \geqslant 1$, we have $\pi(2^k) \leqslant \frac{3}{k} \cdot 2^k = 3\log(2) \cdot 2^k/\log(2^k)$.

We prove this by induction on k. For $n \in \mathbb{N}$, define $N = (2n)! \div (n!)^2$, as in Proposition 4.21, so:

$$2^{2n} \geqslant N \geqslant \prod_{n<p\leqslant 2n} p \geqslant n^{\pi(2n)-\pi(n)} \implies \pi(2n) - \pi(n) \leqslant \frac{n}{\log n} \cdot 2\log 2.$$

Suppose we know that $\pi(2^k) \leqslant \frac{3}{k} \cdot 2^k$. Then take $n = 2^k$ to obtain the bound:

$$\pi(2n) = \pi(2^{k+1}) \leqslant \pi(n) + \frac{n}{\log n} \cdot 2\log 2 = \frac{3}{k} \cdot 2^k + \frac{2^k}{k\log 2} \cdot 2\log 2 = \frac{5 \cdot 2^k}{k}.$$

The induction step is complete for $k \geqslant 5$, and the earlier cases are easy to check. So in fact it must hold for all k. We now try to extend this proof to other integers.

Now, let k be such that $2^k \leqslant X < 2^{k+1}$. We see that $\pi(X) \leqslant \pi(2^{k+1}) \leqslant 6\log(2) \cdot 2^k/\log(2^k)$. But:

$$\frac{d}{dX}\left(\frac{X}{\log X}\right) = \frac{\log(X) - 1}{\log(X)^2} > 0 \text{ where } X > e \implies \frac{X}{\log(X)} \text{ is strictly increasing for } X > 4.$$

So $\pi(X) \leqslant 6\log(2) \cdot 2^k/\log(2^k) \leqslant (6\log 2) \cdot (X/\log X)$, which proves our upper bound with c_2. □

Proof: (Lower bound). Take n and N as before. Then we have:

$$\frac{2^{2n}}{2n} \leqslant N = \prod_{p\leqslant 2n} p^{v_p(N)} = (2n)^{\pi(2n)} \implies \pi(2n) + 1 \geqslant \frac{2n\log(2)}{\log(2n)}.$$

Rearranging yields $\pi(2n) \geqslant \log 2 \cdot 2n/(\log 2n) - 1$. For $X > 4$, choose n with $2n \leqslant X < 2n+2$, so:

$$\pi(X) \geqslant \pi(2n) \geqslant \frac{2n}{\log(2n)} \cdot \log(2) - 1 \implies \pi(X) \geqslant \frac{X-2}{\log(X)} \cdot \log(2) - 1.$$

This is similar to the inequality we wish to prove, but not exactly. We want to find a lower bound of the form specified in the theorem, by proving:

$$\frac{X-2}{\log(X)} \cdot \log(2) - 1 \geqslant \frac{\log(2)}{2} \cdot \frac{X}{\log(X)} \iff \frac{\log(2)}{2} \cdot \frac{X}{\log(X)} - \frac{2\log(2)}{\log(X)} - 1 \geqslant 0.$$

The left hand side of this inequality is increasing for $X > 4$, and in fact is satisfied for $X = 16$, where it is equal to $1/2$.

Therefore we have proved the result for all $X \geqslant 16$, and we may check the remaining values individually. We wish to show that:

$$\pi(X) \geqslant \frac{\log(2)}{2} \cdot \frac{X}{\log(X)} \text{ with } 4 < X \leqslant 16.$$

The right hand side is maximised when $X = 16$, yielding $\frac{1}{2}\log(2) \cdot 4/\log(2) = 2$. However, the left hand side is always at least 2 when $X > 4$, since 2 and 3 are prime! Thus we have shown the lower bound for all $X > 4$, as required.

Thus the lower bound and upper bound both hold, proving Chebyshev's theorem. □

Now, we prove another bound, this time as an auxiliary proposition in our pursuit of a new and exciting result: Bertrand's Postulate.

Proposition 4.23 (Prime Product Bound)

For $X \geqslant 1$, let $P(X)$ be the product of all primes which are at most X. Then $P \leqslant 4^X$.

Proof: It is enough to show this for integer values $X = m$, since $P(X) = P(\lfloor X \rfloor) \leqslant 4^{\lfloor X \rfloor} \leqslant 4^X$. We may check manually that $P(1) = 1 \leqslant 4$, and $P(2) = 2 \leqslant 16$.

We now use strong induction. Suppose $m \geqslant 2$, with $P(k) \leqslant 4^k$ for all $k \leqslant n$. We will show that $P(m+1) \leqslant 4^{m+1}$, which suffices to show the proposition for all $m \in \mathbb{N}$ (and therefore all $X \geqslant 1$).

If m is odd, then $m+1 \geqslant 4$ is even and thus not prime, so $P(m) = P(m+1) \leqslant 4^m \leqslant 4^{m+1}$.

If $m = 2\ell$ is even, then write:

$$P(m+1) = \prod_{p \leqslant \ell+1} p \times \prod_{\ell+2 \leqslant p \leqslant 2\ell+1} p = P(\ell+1) \times \prod_{\ell+2 \leqslant p \leqslant 2\ell+1} p.$$

By considering $N = (2\ell+1)! \div (\ell! \times (\ell+1)!)$, or $2\ell+1$ choose ℓ, we see that the product:

$$\left(\prod_{\ell+2 \leqslant p \leqslant 2\ell+1} p \right) \text{ must be a factor of } \frac{(2\ell+1) \times (2\ell) \times \cdots \times (\ell+2)}{\ell \times (\ell-1) \times \cdots \times 2 \times 1} = N.$$

Similarly, $2^{2\ell+1} = (1+1)^{2\ell+1} \geqslant 2N$, as N appears twice in the binomial sum (as both the ℓ^{th} and $(\ell+1)^{\text{th}}$ term). We therefore have $2 \cdot 2^{2\ell} \geqslant 2N$, so $N \leqslant 4^\ell$.

Combining these results yields $P(m+1) \leqslant P(\ell+1) \times N \leqslant 4^{\ell+1} \times 4^\ell = 4^{m+1}$ as required. □

We are now able to state and prove Bertrand's postulate!

Theorem 4.24 (Bertrand's Postulate)

If $n \in \mathbb{N}$ is greater than 1, there is a prime number p with $n < p < 2n$.

Proof: Assume $n \geqslant 3$ (since $n = 2$ yields $p = 3$), and assume there is no such prime p. As usual, we consider $N = (2n)! \div (n!)^2$.

By assumption, if $p \mid N$ is prime, then $p \leqslant n$, since any prime p greater than $2n$ would not divide N, and we have no primes between n and $2n$.

We now use the third part of Proposition 4.21, which yields $p \leqslant 2n/3$. Consider the factorisation:

$$N = N_1 \times N_2 \text{ where } N_1 = \prod_{p|N:\, v_p(N)=1} p \text{ and } N_2 = \prod_{p|N:\, v_p(N)>1} p^{v_p(N)}.$$

Now, the first sum N_1 is equal to $P(2n/3) \leqslant 4^{2n/3}$, with P as in Proposition 4.23. But if $p^2 \mid N$, then by the fourth part of Proposition 4.21, we have $p^2 \leqslant 2n$, and so $p \leqslant \sqrt{2n}$.

The number of primes in the product defining N_2 is then at most $\sqrt{2n}$, and each is at most $2n$. So in fact $N_2 \leqslant (2n)^{\sqrt{2n}}$. The first part of Proposition 4.21 then gives us the bound:

$$\frac{2^{2n}}{2n} \leqslant N = N_1 \times N_2 \leqslant 2^{4n/3} \times (2n)^{\sqrt{2n}} \implies 2^{2n/3} \leqslant (2n)^{1+\sqrt{2n}}.$$

Taking logarithms yields $\frac{2}{3} n \log(2) \leqslant (1 + \sqrt{2n}) \log(2n)$. But then the left hand side grows linearly in N, while the right hand side grows as the product of a term in $N^{1/2}$ and one in $\log(N)$, which is clearly asymptotically slower! For some value of N, we surely must reach a contradiction.

In fact, for $n \geqslant 468$, this is a contradiction! Thus Bertrand's Postulate is true for all $n \geqslant 468$, and we need only check values below this.

- 479 is prime, so in fact the postulate is true for all $239 \leqslant n < 468$ as well.
- 239 is prime, so in fact the postulate is true for all $120 \leqslant n < 239$ as well.
- 127 is prime, so in fact the postulate is true for all $64 \leqslant n < 120$ as well.
- 67 is prime, so in fact the postulate is true for all $34 \leqslant n < 63$ as well.
- 37 is prime, so in fact the postulate is true for all $19 \leqslant n < 34$ as well.
- 23 is prime, so in fact the postulate is true for all $12 \leqslant n < 19$ as well.
- 13 is prime, so in fact the postulate is true for all $7 \leqslant n < 12$ as well.
- Finally, the primes 3, 5, and 7 take care of the remaining cases 2, 3, 4, 5, and 6.

Thus the proof holds for all $n > 1$, as required! □

Corollary: For all primes p, there is a prime q with $q < p < 2q$.

Note: We could instead have proved this (again for $n \geqslant 468$), then exhibited the prime sequence 2, 3, 5, 7, 13, 23, 43, 83, 163, 317, and 631 to prove the result for all $p > 1$.

5 Continued Fractions

It is easy to express many numbers in decimal form. These are convenient especially for rational numbers, which have either terminating or eventually periodic decimal expansions, and they are convenient for arithmetic and comparing numbers.

Now, we look at a different way of representing real numbers: continued fractions. These do not have the same properties. They are difficult to perform arithmetic with, but they excel at enabling us to find good rational approximations to real numbers.

Example 5.1 (Approximating π)

One idea for finding a nice approximation to π is to truncate its decimal representation at some point and take the rational number implied by that. For example, $\pi \approx 3.14159$, so:

$$\left|\pi - \frac{314159}{100000}\right| \approx \frac{1}{376848} < 3 \times 10^{-6}.$$

But this is in some ways a wasteful approximation. Consider the approximation:

$$\left|\pi - \frac{355}{113}\right| \approx \frac{1}{3748629} < 3 \times 10^{-7}.$$

This uses much smaller numbers to approximate π, but is almost ten times as good! This approximation can be found using the *continued fraction decomposition* of π, which the rest of this section is devoted to investigating.

In fact, continued fractions generate the *best* rational approximations possible!

Definition 5.2 (Continued Fraction)

Take a sequence of real numbers $a_0, a_1, a_2, a_3, \dots a_n$ with $a_i > 0$ for all $i > 0$. Then we define the *continued fraction* to be:

$$[a_0, a_1, a_2, \dots, a_n] = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}}$$

terminating at a_n. Explicitly, we define the two-term continued fraction to be:

$$[a_0, a_1] = a_0 + \frac{1}{a_1}$$

and in general, use the recursive definition $[a_0, a_1, a_2, , \dots, a_n] = [a_0, a_1, \dots, a_{n-2}, [a_{n-1}, a_n]]$.

Now, we claim that we can assign any real number a continued fraction! We do so using something called the *continued fraction algorithm*, which takes as input $\theta \in \mathbb{R}$, and returns as output two sequences $\theta_0, \theta_1, \theta_2, \cdots \in \mathbb{R}$ and $a_0, a_1, a_2, \cdots \in \mathbb{Z}$ such that:

1. for all $i \geqslant 1$, we have $\theta_i > 1$ and $a_i \geqslant 1$.
2. for all $n \geqslant 0$ where θ_{n+1} is defined, we have $\theta = [a_0, a_1, \dots, a_n, \theta_{n+1}]$.

How does the algorithm work? We prepare by setting $\theta_0 = \theta$ and $a_0 = \lfloor \theta_0 \rfloor$. If $a_0 = \theta_0 = \theta$, then θ was in fact an integer to begin with, and we may stop. Otherwise, $0 < \theta_0 - a_0 < 1$, and so we may define the reciprocal to be $\theta_1 = (\theta_0 - a_0)^{-1} > 1$, so that $\theta = [a_0, \theta_1]$ as required.

Now, we set $a_1 = \lfloor \theta_1 \rfloor$. If $a_1 = \theta_1$, then stop. Otherwise, set $\theta_2 = (\theta_1 - a_1)^{-1} > 1$, so $\theta = [a_0, a_1, \theta_2]$. The algorithm continues like this for all n (until we possibly stop).

There are thus two possibilities. Either we stop, so that $\theta = [a_0, a_1, \dots a_n]$ for some sequence of $n+1$ integers, or we continue forever.

Note: If we do stop, then in particular θ must be rational, as we have expressed it as some finite continued fraction, and we can multiply through to find a single fraction.

In the other case, the sequences of θ_i and a_i are both infinite. We write $\theta = [a_0, a_1, a_2, \dots]$. In both cases, we call this expression the *continued fraction expansion* (CFE) of θ.

Note: Of course, we have not yet made precise the notion of the infinite continued fraction. We do this later on, though it is easy to show that the truncations of an infinite CFE converge.

Example 5.3 (Continued Fraction Expansion)

Take $\theta_0 = \theta = 59/13$, a rational number. What is the CFE of θ?

1. We compute $a_0 = \lfloor 59/13 \rfloor = 4$, so $\theta_1 = (7/13)^{-1} = 13/7$.
2. We compute $a_1 = \lfloor 13/7 \rfloor = 1$, so $\theta_2 = (6/7)^{-1} = 7/6$.
3. We compute $a_2 = \lfloor 7/6 \rfloor = 1$, so $\theta_3 = (1/6)^{-1} = 6$.
4. We compute $a_3 = \lfloor 6 \rfloor = 6 = \theta_3$, so we stop here as we have found an integer.

Therefore we write $59/13 = [a_0, a_1, a_2, a_3] = [4, 1, 1, 6]$. Alternatively, this is:

$$59/13 = [4, 1, 1, 6] = 4 + \frac{1}{1 + \frac{1}{1 + \frac{1}{6}}}.$$

Clearly, any θ with a finite continued fraction is rational. Here, we found the converse: we started with some rational θ, and indeed the algorithm terminated and returned a finite continued fraction. We might conjecture that this always happens: any finite θ causes the algorithm to terminate.

Proposition 5.4 (Rational Continued Fractions)

Let $\theta \in \mathbb{R}$. Then the continued fraction of θ is finite if and only if $\theta \in \mathbb{Q}$.

Proof: The first direction is easy: we can use the continued fraction to generate a fraction equal to θ, which means θ must therefore be rational.

If $\theta \in \mathbb{Z}$, then $\theta = [a_0]$. Suppose $\theta \in \mathbb{Z}$ and $\theta_1 = r_1/r_2$, where $r_1 > r_2 > 0$ are coprime integers. We apply the Euclidean Algorithm (1.4) to r_1 and r_2, which generates a sequence with:

$$\begin{aligned} r_1 &= q_1 r_2 + r_3 & 0 < r_3 < r_2 \\ r_2 &= q_2 r_3 + r_4 & 0 < r_4 < r_3 \\ &\dots & \dots \\ r_n &= q_n r_{n+1} + r_{n+2} & 0 < r_{n+2} < r_{n+1} \\ r_{n+1} &= q_{n+1} r_{n+2} & 1 = r_{n+2} \end{aligned}$$

where the last equality is because r_{n+1} and r_{n+2} are coprime. We now claim that in fact for *all* i, we have $\theta_i = r_i/r_{i+1}$ (until $i = n+1$, of course). This is true for $i = 1$ by definition of r_1 and r_2.

Suppose $\theta_i = r_i/r_{i+1}$. Then $r_i = q_i r_{i+1} + r_{i+2}$, so $r_i/r_{i+1} = \theta_i = q_i + r_{i+2}/r_{i+1}$, where $r_{i+2} < r_{i+1}$ and q_i is an integer. Then by construction, we have $a_i = q_i$, and so $\theta_{i+1} = (r_{i+2}/r_{i+1})^{-1}$, which is exactly what we require to prove the claim.

In particular, $\theta_{n+1} = r_{n+1}/r_{n+2} = q_{n+1} \in \mathbb{Z}$, so we terminate in $n+1$ steps. □

Corollary: This proof shows that the q_i are in fact the terms in the continued fraction expansion, which we wrote as the a_i. In general, we call them the *partial quotients* of θ.

Definition 5.5 (Convergents)

Suppose $a_0, a_1, a_2, \cdots \in \mathbb{Z}$, with $a_i \geqslant 1$ for all $i \geqslant 1$. Then define two sequences (p_n) and (q_n) recursively. Define $p_0 = a_0$ and $q_1 = 1$, and define $p_1 = a_0a_1 + 1$ and $q_1 = a_1$. Then:

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

Alternatively, one could take $p_{-1} = 1$ and $q_{-1} = 0$ to extend this recursion. The q_n must all be positive (since they do not use a_0), and moreover (q_n) must be an increasing sequence.

This definition can be written in matrix form:

$$\begin{pmatrix} p_{n+1} & p_n \\ q_{n+1} & q_n \end{pmatrix} = \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} \begin{pmatrix} a_{n+1} & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_{n+1} & 1 \\ 1 & 0 \end{pmatrix}.$$

If $\theta \in \mathbb{R}$ has continued fraction expansion $\theta = [a_0, a_1, a_2, \dots]$, then the sequence given by the ratios (p_n/q_n) for $n \geqslant 0$ are called the *convergents* of θ.

Why are these useful? In fact, they are what give us good rational approximations to θ.

Proposition 5.6 (Convergents)

Suppose that we have $a_0, a_1, a_2, \dots$ as usual. Then $p_n/q_n = [a_0, a_1, \dots a_n]$. Moreover, for all $n \geqslant 1$, we have $p_nq_{n-1} - q_np_{n-1} = (-1)^{n+1}$, and p_n and q_n are coprime.

Now suppose $\beta \in \mathbb{R}$ with $\beta > 0$. Then for all $n \geqslant 1$, we have:

$$\frac{\beta p_n + p_{n-1}}{\beta q_n + q_{n-1}} = [a_0, a_1, \dots, a_n, \beta],$$

and this is a real number strictly between p_n/q_n and p_{n-1}/q_{n-1}.

Proof: In fact, choosing $\beta = a_{n+1}$ makes the first part follow from the equality. Similarly, we may take determinants in the matrix equation for p_n and q_n:

$$\underbrace{\det \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix}}_{p_nq_{n-1} - q_np_{n-1}} = \prod_{i=0}^{n} \underbrace{\det \begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix}}_{-1} = (-1)^{n+1}.$$

This proves that p_n and q_n are coprime, by Bézout's Identity (a corollary to Proposition 1.3). We can also divide by $q_{n-1}q_n$ to find the other identity. It just remains to prove the equality for all positive β, and show that this lies in the claimed range.

We do this by induction on n, using $p_{-1} = 1$ and $q_{-1} = 0$. In the base case with $n = 0$, we have:

$$\frac{\beta a_0 + 1}{\beta} = a_0 + \frac{1}{\beta} = [a_0, \beta].$$

Now suppose this is true for n. We compute $[a_0, \dots, a_n, a_{n+1}, \beta] = [a_0, \dots, a_n, [a_{n+1}, \beta]]$. Define $\gamma = [a_{n+1}, \beta] = a_{n+1} + 1/\beta$. By induction, we have:

$$[a_0, \dots, a_n, \gamma] = \frac{\gamma p_n + p_{n-1}}{\gamma q_n + q_{n-1}} = \frac{(a_{n+1} + 1/\beta)p_n + p_{n-1}}{(a_{n+1} + 1/\beta)q_n + q_{n-1}} = \frac{a_{n+1}\beta p_n + p_n + p_{n-1}}{a_{n+1}\beta q_n + q_n + q_{n-1}} = \frac{\beta p_{n+1} + p_n}{\beta q_{n+1} + q_n}.$$

It remains to show that this is between p_n/q_n and p_{n-1}/q_{n-1}. We know that the absolute difference between these two numbers is $1/q_nq_{n-1}$. We now use the fact that if $x/y > x'/y'$ for positive y and y', then $x/y > (x + x')/(y + y') > x'/y'$. Taking x/y and x'/y' to be the larger and smaller of the two convergents yields the result directly. □

Corollary: In the important special case where $\theta = [a_0, a_1, \dots]$, we have $\theta = [a_0, a_1, \dots, a_n, \theta_{n+1}]$. This gives us the equation:

$$\theta = \frac{\theta_{n+1} p_n + p_{n-1}}{\theta_{n+1} q_n + q_{n-1}} \quad \text{for all } n.$$

Theorem 5.7 (Irrational Continued Fractions)

Let $\theta \in \mathbb{R} \setminus \mathbb{Q}$. Then for all $n \geqslant 0$, θ lies strictly between p_n/q_n and p_{n+1}/q_{n+1}, and the error term given by $|\theta - p_n/q_n| < 1/q_n q_{n+1}$. Moreover, the convergents $p_n/q_n \to \theta$ as $n \to \infty$.

Proof: By the above corollary (applied to θ_{n+2}), θ clearly lies between the two convergents. Since θ is irrational, this must be strictly true, as it cannot be equal to the endpoints.

This interval has length $1/q_n q_{n+1}$, and this bounds error term.

The convergence of the convergents follows straightforwardly by the fact that $q_n \to \infty$, since they form a strictly increasing sequence of integers. □

Corollary: θ is determined entirely by its continued fraction expansion.

Corollary: Without too much more work, one can show that the map between $\mathbb{R}$ and the set of integer sequences which meet the condition for a continued fraction, which sends each $\theta \in \mathbb{R}$ to its continued fraction as a sequence, is in fact a bijection.

Note: This justifies the truncation of infinite continued fractions for irrational numbers! For such a sequence, θ is equal to the limit of the continued fractions truncated at each point.

Recall from Example 5.1 that $\pi \approx 355/113$, and that this is a very good approximation. This is in fact one of the convergents for π. The CFE of π is infinite, of course, but begins $[3, 7, 15, 1, 292, 1]$. The first few convergents are therefore:

$$[3] = 3, \quad [3, 7] = 3 + 1/7 = 22/7, \quad [3, 7, 15] = 3 + 15/106 = 333/106, \quad [3, 7, 15, 1] = 355/113.$$

These are all "unusually good" approximations for π, given the small size of their denominators. In fact, they are the *best* possible approximations, and we can make this notion precise!

Theorem 5.8 (Rational Approximation Theorem)

Suppose $\theta \in \mathbb{R} \setminus \mathbb{Q}$, and let $p, q \in \mathbb{Z}$ with $q > 0$. Then:

1. If $q < q_{n+1}$, then $|q\theta - p| \geqslant |q_n \theta - p_n|$.
2. If $q \leqslant q_n$, then $|\theta - p/q| \geqslant |\theta - p_n/q_n|$.

That is, any fraction with a denominator which is smaller than q_n cannot be a strictly better approximation to θ than p_n/q_n is.

Proof: Clearly, if (1) holds, then $|\theta - p/q| = 1/q\,|q\theta - p| \geqslant 1/q_n\,|q_n\theta - p_n| = |\theta - p_n/q_n|$.

To show (1), consider the matrix given in Definition 5.5, with determinant $(-1)^n$. We can thus find integers u and v with:

$$\begin{pmatrix} p_{n+1} & p_n \\ q_{n+1} & q_n \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}.$$

This means $q\theta - p = u(q_{n+1}\theta - p_{n+1}) + v(q_n\theta - p_n)$. If $u = 0$, then v is a non-zero integer (if not, then $q\theta - p = 0$, which contradicts the irrationality of θ), and so $|q\theta - p| \geqslant |q_n\theta - p_n|$.

Suppose $u \neq 0$. As $uq_{n+1} + vq_n = q$ and $q < q_{n+1}$, $v \neq 0$ too, and also u and v must have opposite signs. Also, $q_n\theta - p_n$ and $q_{n+1}\theta - p_{n+1}$ have opposite signs, as θ lies between the two convergents. Thus the two products have the same sign, so the absolute value of $u(q_{n+1}\theta - p_{n+1}) + v(q_n\theta - p_n)$ is $|q\theta - p| = |u|\,|q_{n+1}\theta - p_{n+1}| + |v|\,|q_n\theta - p_n| \geqslant |q_n\theta - p_n|$ as required, since $|v| \geqslant 1$. □

So we know that the continued fraction convergents are the best approximations of θ, in a sense. How good are they in absolute terms?

Theorem 5.9 (Convergent Error Bound)

Suppose $\theta \in \mathbb{R} \setminus \mathbb{Q}$. Then for all $n \geqslant 1$, at least one of the two convergents $p/q = p_n/q_n$ or p_{n+1}/q_{n+1} satisfies $|\theta - p/q| < 1/2q^2$.

In fact, if p/q is a fraction with $|\theta - p/q| < 1/2q^2$, then p/q is a convergent of θ. That is, the *only* fractions which are "this accurate" are those generated by the continued fraction of θ.

Proof: We know that $\theta - p_n/q_n$ and $\theta - p_{n+1}/q_{n+1}$ have opposite signs. This means that the sum of the absolute values is in fact $|p_n/q_n - p_{n+1}/q_{n+1}| = 1/q_nq_{n+1}$. The AM-GM inequality yields:

$$|\theta - p_n/q_n| + |\theta - p_{n+1}/q_{n+1}| = \left|\frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}}\right| = \frac{1}{q_nq_{n+1}} < \frac{1}{2}\left(\frac{1}{q_n^2} + \frac{1}{q_{n+1}^2}\right).$$

But then the first part of the theorem certainly holds: if not, we would have a contradiction.

Now suppose $|\theta - p/q| < 1/2q^2$. Choose the $n \geqslant 1$ such that $q_n \leqslant q < q_{n+1}$. Then the Rational Approximation Theorem (5.8) gives us the inequality $|q\theta - p| \geqslant |q_n\theta - p_n|$. Now, use:

$$\left|\frac{p}{q} - \frac{p_n}{q_n}\right| \leqslant \left|\frac{p}{q} - \theta\right| + \left|\theta - \frac{p_n}{q_n}\right| = \frac{|q\theta - p|}{q} + \frac{|q_n\theta - p_n|}{q_n} \leqslant \left(\frac{1}{q} + \frac{1}{q_n}\right)|q\theta - p| < \left(\frac{1}{2q^2} + \frac{1}{2qq_n}\right)$$

from the triangle inequality. Suppose that $p/q \neq p_n/q_n$. Then $|p/q - p_n/q_n| \geqslant 1/qq_n$, since this is a rational number with denominator dividing qq_n. Combining these results yields:

$$\frac{1}{qq_n} < \left(\frac{1}{q} + \frac{1}{q_n}\right) \cdot \frac{1}{2q} \implies 2q < q_n + q.$$

But this is a contradiction, since we chose n such that $q \geqslant q_n$. This means that $p/q = p_n/q_n$, and so any rational approximation to θ with this level of accuracy must be a convergent. □

Note: It is *not* the case that only one convergent is an integer! Of course $[a_0] = a_0$ is an integer, but it is possible that $[a_0, a_1] = a_0 + 1$ if $a_1 = 1$: this happens for e, for instance. Since the CFE of e begins $[2, 1, \dots]$, the convergents begin 2, 3, and so on. In fact, this is true of all numbers of the form $n + \alpha$, where $n \in \mathbb{Z}$ and $1/2 < \alpha < 1$.

5.1 Pell's Equation

We now apply this theory to find solutions to *Pell's Equation*.

Definition 5.10 (Pell's Equation)

For $d \in \mathbb{N}$ not a square number, *Pell's equation* is $X^2 - dY^2 = 1$.

We can use continued fractions to find solutions beyond the trivial solution $(X, Y) = (1, 0)$. If $(p, q) \in \mathbb{N}^2$ is a solution, then we can complete the square to find:

$$(p - q\sqrt{d})(p + q\sqrt{d}) = p^2 - dq^2 = 1 \implies p/q - \sqrt{d} = \frac{1}{q^2} \cdot \frac{1}{p/q + \sqrt{d}}.$$

This is a positive number, and so $p/q > \sqrt{d}$. Moreover, the absolute value of their difference is less than $1/2q^2$, because $p/q + \sqrt{d} > 2\sqrt{d} > 2$. This means that p/q is a convergent of $\sqrt{d}$.

We thus want to study the CFE of numbers of the form $r + s\sqrt{d}$, where $r, s \in \mathbb{Q}$ and $s \neq 0$. What do these look like, and do they have any special properties? Studying numbers like this will allow us to find solutions to Pell's equation for d.

Example 5.11 (CFE of $\sqrt{6}$)

Take $\theta_0 = \theta = \sqrt{6}$, an irrational number. What is the CFE of θ?

1. We compute $a_0 = \lfloor \sqrt{6} \rfloor = 2$, so $\theta_1 = (\sqrt{6} - 2)^{-1} \approx 2.225$.
2. We compute $a_1 = \lfloor \theta_1 \rfloor = 2$, so $\theta_2 = (0.225)^{-1} \approx 4.449$.
3. We compute $a_2 = \lfloor \theta_2 \rfloor = 4$, so $\theta_3 = (0.449)^{-1} \approx 2.225 = \theta_1$.

In fact, we can see where this goes! The continued fraction algorithm will repeat from here, since the value of each θ only depends on the previous one.

Thus $\theta = [2, \theta_1] = [2, 2, \theta_2] = [2, 2, 4, \theta_3] = [2, 2, 4, \theta_1] = [2, 2, 4, 2, 4, 2, 4, 2, 4, \dots]$.

This is a lot like the decimal expansions of rational numbers becoming eventually periodic!

Definition 5.12 (Periodic)

Suppose (a_n) is a sequence of integers, all of which are positive except possibly a_0. Then we say the continued fraction $[a_0, a_1, a_2, \dots]$ is *essentially periodic* if there are $m \geqslant 0$ and $k \geqslant 1$ such that for all $n \geqslant m$, we have $a_n = a_{n+k}$.

We say that it is *purely periodic* if this still holds for $m = 0$.

For such a continued fraction, we write $[a_0, a_1, \dots, a_{m-1}, \overline{a_m, a_{m+1}, a_{m+2}, \dots, a_{m+k-1}}]$

Theorem 5.13 (Lagrange's Continued Fraction Theorem)

Suppose $\theta \in \mathbb{R} \setminus \mathbb{Q}$. Then the CFE of θ is essentially periodic if and only if θ is a *quadratic irrational*: that is, $\theta = r + s\sqrt{d}$ for rational r and s and non-square $d \in \mathbb{N}$.

Proof: If $\theta \in \mathbb{R} \setminus \mathbb{Q}$, then θ is a quadratic irrational if and only if it satisfies some equation of the form $a\theta^2 + b\theta + c$, where a, b, c are integers with $a \neq 0$.

Suppose we have a purely periodic CFE for θ. Then we can write:

$$\theta = [a_0, \dots, a_n, \overline{a_0, \dots, a_n}] = [a_0, \dots, a_n, \theta] = \frac{p_n\theta + p_{n-1}}{q_n\theta + q_{n-1}}.$$

Multiplying through by the denominator yields $q_n\theta^2 + (q_{n-1} - p_n)\theta - p_{n-1} = 0$, so indeed θ is a quadratic irrational, since $q_n \neq 0$. If instead θ has an esentially periodic CFE, then:

$$\theta = [a_0, \dots, a_{m-1}, \overline{a_m, \dots, a_{m+k-1}}] = [a_0, \dots, a_{m-1}, \sigma].$$

But then σ is a quadratic irrational, so $\sigma = r + s\sqrt{d}$. But then θ is a finite nested fraction involving σ, and so it is easy to rewrite it as $r' + s'\sqrt{d}$.

Conversely, suppose that $\theta \in \mathbb{R} \setminus \mathbb{Q}$ is a solution to $a\theta^2 + b\theta + c$. Then $f(x, y) = ax^2 + bxy + cy^2$ is a binary quadratic form with integer coefficients, with the property that $f(\theta, 1) = 0$. For any $n \geqslant 1$, associate another binary quadratic form $f_n(x, y) = f(p_n x + p_{n-1}y, q_n x + q_{n-1}y)$. Now:

$$\theta = [a_0, \dots, a_n, \theta_{n+1}] = \frac{p_n\theta_{n+1} + p_{n-1}}{q_n\theta_{n+1} + q_{n-1}} \implies f_n(\theta_{n+1}, 1) = (q_n\theta_{n+1} + q_{n-1})^2 f(\theta, 1) = 0.$$

We now claim that there are only finitely many possibilities for f_n, which means there are only finitely many possibilities for θ_{n+1} as n varies. By the pigeonhole principle, there is thus a repeated $\theta_{n+1} = \theta_{n+1+k}$, but then as in Example 5.11, the continued fraction much henceforth repeat.

Therefore if the claim is true, there is an eventually periodic continued fraction for θ, which proves the theorem. Why is the claim true?

Write $f_n(x,y) = A_n x^2 + B_n xy + C_n y^2$. Then $A_n = f_n(1,0) = f(p_n, q_n)$, and $C_n = f_n(0,1) = A_{n-1}$. Now, observe that $B_n^2 - 4A_nC_n = \operatorname{disc} f_n = \operatorname{disc} f \cdot (p_nq_{n-1} - p_{n-1}q_n)^2 = \operatorname{disc} f$.

Since $A_{n-1} = C_n$, and B_n is determined (up to sign) by A_n and C_n, it suffices to show that A_n can take only finitely many values as n varies, and in fact since $A_n \in \mathbb{Z}$ we may merely show that it is bounded. We factor $f(x,1) = a(x-\theta)(x-\theta')$.

Then $|A_n| = |f(p_n,q_n)| = q_n^2 \,|a|\, |\theta - p_n/q_n|\, |\theta' - p_n/q_n|$. But we know that $|\theta - p_n/q_n|$ is bounded above by $1/q_nq_{n+1}$, which means that we have the bound:

$$|A_n| \leqslant \frac{q_n\,|a|}{q_{n+1}} \times \left|\frac{p_n}{q_n} - \theta'\right| \leqslant |a| \times \left|\frac{p_n}{q_n} - \theta'\right|.$$

Since the sequence of convergents tends to θ', the sequence of $|p_n/q_n - \theta'|$ tends to $|\theta - \theta'|$ as n grows. In particular, it is bounded as n varies!

But then the values A_n can take are bounded, and therefore finite, and therefore there are finitely many f_n, and therefore finitely many θ_{n+1}, and therefore one repeats, and therefore the pattern repeats, and therefore θ has an essentially periodic CFE. □

Theorem 5.14 (Galois Continued Fraction Theorem)

Suppose that $\theta = r + s\sqrt{d}$ is a quadratic irrational. Then the CFE of θ is purely periodic if and only if $\theta > 1$ and $-1/\theta' > 1$, where $\theta' = r - \sqrt{d}$ is the *conjugate irrational* of θ. Moreover, these conditions are symmetric in θ and $-1/\theta'$, so $-1/\theta'$ will also have a continued fraction expansion which is purely periodic. If this is the case, then:

$$\theta = [\overline{a_0, a_1, \ldots, a_{n-1}, a_n}] \implies -1/\theta' = [\overline{a_n, a_{n-1}, \ldots, a_1, a_0}].$$

Proof: Omitted. □

We now apply these results to the problem of solving Pell's Equation (5.10). Unfortunately, since $\sqrt{d} > 1$, if $\theta = \sqrt{d}$, then $-1/\theta' = 1/\sqrt{d} < 1$, so we cannot apply Theorem 5.13. However, we may instead take $\theta = \theta_0 = \sqrt{d}$, $a_0 = \lfloor \theta \rfloor$, and let $\theta_1 = (\theta_0 - a_0)^{-1} > 1$. Then:

$$\frac{-1}{\theta_1'} = \frac{-1}{(-\sqrt{d} - a_0)^{-1}} = \sqrt{d} + a_0 > 1.$$

Thus θ_1 has a purely periodic CFE, by the above theorem, and so the CFE of $\sqrt{d}$ becomes periodic after just one step a_0. For example, $\sqrt{6} = [2, \overline{2, 4}]$ as in Example 5.11.

Theorem 5.15 (Pell's Theorem)

Let $d \in \mathbb{N}$ be a non-square number. Then Pell's equation $X^2 - dY^2 = 1$ has integer solutions.

Proof: Suppose $\sqrt{d} = [a_0, \overline{a_1, a_2, \ldots, a_n}] = [a_0, a_1, a_2, \ldots, a_n, \theta_1]$. Then we can write:

$$\sqrt{d} = \frac{p_n\theta_1 + p_{n-1}}{q_n\theta_1 + q_{n-1}} = \frac{(p_n - a_0p_{n-1}) + p_{n-1}\sqrt{d}}{(q_n - a_0q_{n-1}) + q_{n-1}\sqrt{d}}$$

Multiplying through yields $dq_{n-1} + (q_n - a_0q_{n-1})\sqrt{d} = (p_n - a_0p_{n-1}) + p_{n-1}\sqrt{d}$. Now, we can equate the integer part and coefficients of $\sqrt{d}$, so $dq_{n-1} = p_n - a_0p_{n-1}$ and $q_n - a_0q_{n-1} = p_{n-1}$.

We can now take n to be even without loss of generality, since otherwise we can take the CFE to have a period of $2n$. Then evaluating $p_{n-1}^2 - dq_{n-1}^2$ using the above identity yields:

$$p_{n-1}^2 - dq_{n-1}^2 = p_{n-1}(q_n - a_0q_{n-1}) - q_{n-1}(p_n - a_0p_{n-1}) = p_{n-1}q_n - p_nq_{n-1} = (-1)^n = 1,$$

by Proposition 5.6. This means (p_{n-1}, q_{n-1}) is a valid integer solution to Pell's equation. □

Corollary: In fact, since we can force the solution to have any even multiple of the original period and still generate a valid proof, there are infinitely many solutions to Pell's equation!

Corollary: If $d \in \mathbb{N}$ is not square, then the set of solutions (p, q) to $p^2 - dq^2 = \pm 1$ are in fact the convergents (p_{kn-1}, q_{kn-1}), where k is the minimal period of the CFE of $\sqrt{d}$.

Example 5.16 (Solving Pell's Equation)

In Example 5.11, we found $\sqrt{6} = [2, \overline{2, 4}]$. We want to use this result to find integer solutions to Pell's equation with $d = 6$, namely $X^2 - 6Y^2 = 1$.

Here, the period is 2: we have $a_1 = 2$ and $a_2 = 4$. Since 2 is even, we expect to find a solution given by $n = 2$, namely (p_1, q_1).

We can find these using $p_1/q_1 = [2, 2] = 2 + \frac{1}{2} = \frac{5}{2}$, so $p_1 = 5$ and $q_1 = 2$. Indeed, we see that:

$$5^2 - 6 \cdot 2^2 = 25 - 24 = 1$$

and so this is indeed a solution to Pell's equation!

What about $d = 17$? In fact, the CFE is very easy to find here. Clearly, $4 < \sqrt{17} < 5$, so we take $a_0 = 4$. Then, $\theta_1 = (\sqrt{17} - 4)^{-1} = \sqrt{17} + 4 = 8 + (\sqrt{17} - 4)$, so in fact $\theta_n = \theta_1$ for all n. This gives us a CFE of $\sqrt{17} = [4, \overline{8}]$.

If we want to solve $X^2 - dY^2 = -1$, $(p_0, q_0) = (4, 1)$ should work: indeed, $16 - 17 = -1$. To solve Pell's equation, we must take the next convergent (and in fact all odd convergents).

Here, $p_1/q_1 = [4, 8] = 4 + \frac{1}{8} = 33/8$, so $(p_1, q_1) = (33, 8)$. As desired, we obtain:

$$33^2 - 17 \cdot 8^2 = 1089 - 17 \cdot 64 = 1089 - 1088 = 1.$$

6 Primality Testing and Factorisation

Suppose N is a very large natural number. Can we easily check whether N is prime? In the case where N is not prime, can we find a non-trivial factor?

Ideally, we want algorithms to answer these questions which always return answers in polynomial time. Unfortunately, this is not known to be possible, and in fact it is strongly suspected that it is impossible to create such algorithms. However, we can certainly do better than naïve algorithms!

Note: In fact, the security of the RSA encryption scheme relies on the assumption that large numbers cannot be factorised quickly: further discussion appears in *II Coding and Cryptography*.

6.1 Probabilistic Primality Tests

Note: Here, we usually restrict our analysis to when N is odd. If N is even, this is of course very easy to check, and so N cannot be prime!

There are *probabilistically* polynomial-time algorithms to test primality of numbers. These usually come from necessary (but not sufficient) conditions for numbers to be prime.

Example 6.1 (Fermat's Little Theorem Test)

If p is a prime, then any $1 < a < p$ is coprime to p. Moreover, Fermat's Little Theorem (1.12) yields that $a^{p-1} \equiv 1 \pmod{p}$.

We can show that 15 is not prime. Take $a = 2$, and notice that $2^4 = 16 \equiv 1 \pmod{15}$, so we have $2^{14} \equiv (2^4)^3 \cdot 2^2 \equiv 1^3 \cdot 2^2 \equiv 2^2 \equiv 4 \not\equiv 1 \pmod{15}$, and so 15 cannot be prime!

In general, we can pick random values of a, and test (using Euclid's Algorithm from 1.4) that a is coprime to N. Then, we can compute $a^{N-1} \pmod{N}$: if this is not congruent to 1, then N certainly cannot be a prime number.

However, this test passing is not a sufficient condition for N to be prime! For example, we can compute $3^{90} \equiv 1 \pmod{91}$, but $91 = 7 \times 13$ is not prime.

We can generalise our description of this situation, where a composite number passes this test.

Definition 6.2 (Fermat Pseudoprime)

Let $N \in \mathbb{N}$ be an odd composite number, and let $b \in \mathbb{Z}$ be coprime to N. Then we say that N is a *Fermat pseudoprime to the base* b if $b^{N-1} \equiv 1 \pmod{N}$.

In some sense, N "looks like" a prime number, at least with regard to this test.

Proposition 6.3 (Fermat Pseudoprimes)

If $N \in \mathbb{N}$ is an odd composite number, then:

(a) If $(b, N) = 1$, then whether or not N is a Fermat pseudoprime to the base b depends only on the reduction of b modulo N: that is, the image of b in $\mathbb{Z}/N\mathbb{Z}$.

(b) The subset $B \subseteq (\mathbb{Z}/N\mathbb{Z})^\times$ of bases b to which N is a Fermat pseudoprime is a subgroup.

(c) If there exists a $b_0 \in (\mathbb{Z}/N\mathbb{Z})^\times$ such that N is not a Fermat pseudoprime to the base b_0 (a witness to N being composite), then at least half the bases have this property.

Proof: (a) This is obvious, since $b^{N-1} \pmod{N}$ only depends on $b \pmod{N}$.

(b) We need to show that $1 \in B$, and that B is closed under multiplication. But these are both easy: $1^{N-1} \equiv 1 \pmod{N}$, and if this is true for b and c, then:

$$b^{N-1} \equiv c^{N-1} \equiv 1 \pmod{N} \implies (bc)^{N-1} \equiv b^{N-1} \cdot c^{N-1} \equiv 1 \cdot 1 \equiv 1 \pmod{N}.$$

Thus we have identity and closure under multiplication, so B is indeed a subgroup.

(c) Write $G = (\mathbb{Z}/N\mathbb{Z})^\times$. Then by Lagrange's Theorem from *IA Group Theory*, $|B|$ is a factor of $|G|$. But if there is such a b_0, then B is a non-trivial subgroup and $B \neq G$. This means that we must have $|B| \leqslant \frac{1}{2}|G|$, and so at least half the elements of G are not in B.

Equivalently, at least half the bases are witnesses to N being composite, as required. □

Why is this last property useful? Well, we want to find a primality test for N, and we have shown that Fermat's Little Theorem (as a necessary condition for primality) gives us an easy way to show that N is composite (by finding a witness). We have thus shown that if there is such a witness, in fact at least half the possible numbers we could have tried are witnesses!

The upshot of this is that we can keep trying numbers at random: if there is such a witness, the probability we will not find one within k attempts is at most $1/2^k$.

Note: We might conjecture that there is always at least one witness, and so we can always quickly find them. Unfortunately, this is not true: there are composite numbers N with $B = G$, so that N is a Fermat pseudoprime to *any* base!

Definition 6.4 (Carmichael Number)

We call an odd composite integer N a *Carmichael number* if it is a Fermat pseudoprime to any base $b \in (\mathbb{Z}/N\mathbb{Z})^\times$. The smallest such number is $561 = 3 \times 11 \times 17$.

Note: Robert Daniel Carmichael described the existence of these numbers in 1910, and found the first fifteen of them. He also conjectured that there are infinitely many Carmichael numbers, but this was not proved until the 1990s.

Let's consider another type of pseudoprime.

Definition 6.5 (Euler Pseudoprime)

Let $N \in \mathbb{N}$ be an odd composite number, and let $b \in \mathbb{Z}$ be coprime to N. Then we say that N is an *Euler pseudoprime to the base* b if:

$$b^{(N-1)/2} \equiv \left(\frac{b}{N}\right) \pmod{N}$$

where the right hand side is the Jacobi symbol (2.13).

Corollary: Since $(b, N) = 1$, neither of the sides is zero. We can thus square both sides to obtain $b^{N-1} \equiv (\pm 1)^2 \equiv 1 \pmod{N}$, so any Euler pseudoprime is also a Fermat pseudoprime.

Corollary: For the same reasons, the properties of Fermat pseudoprimes in Proposition 6.3 still hold. This only depends on b modulo N, the set of such bases forms a subgroup of $(\mathbb{Z}/N\mathbb{Z})^\times$, and if not every base works, then in fact at most half the bases work. The second of these is because the Jacobi symbol is multiplicative, which was proved in Proposition 2.14.

What is the advantage of this definition? Are there any numbers like Carmichael numbers for this new definition of a pseudoprime? In fact, there aren't! This allows us to test for counterexamples at random a lot more efficiently, since at least one witness exists (and therefore at least half of the possible bases are witnesses).

Proposition 6.6 (Euler Pseudoprime Effectiveness)

Let N be odd and composite. Then there exists some $b_0 \in (\mathbb{Z}/N\mathbb{Z})^\times$ such that N is not an Euler pseudoprime to the base b_0.

Proof: We split this proof into two cases. Firstly, suppose N is square-free, and write $N = pM$ such that $p \geqslant 3$ is prime and $p \nmid M$. There exists some u such that the Jacobi symbol of u on p is -1. The Chinese Remainder Theorem (1.14) allows us to choose some $b \in \mathbb{Z}$ with $b \equiv u \pmod{p}$ and $b \equiv 1 \pmod{M}$. We claim that b is our witness.

The Jacobi symbol of b on N is:

$$\left(\frac{b}{N}\right) = \left(\frac{b}{p}\right) \cdot \left(\frac{b}{M}\right) = \left(\frac{u}{p}\right) \cdot \left(\frac{1}{M}\right) = -1 \cdot 1 = -1.$$

Suppose now that b was not a witness, and so N were an Euler pseudoprime to the base b. Then we would have $b^{(N-1)/2} \equiv -1 \pmod{N}$. But then this congruence would also hold modulo M, and we know that $b \equiv 1 \pmod{M}$ by construction, which is obviously a contradiction as $M \neq 2$. Thus this b is a witness, which completes the proof in the first case.

If N is not square-free, take a prime p such that $p^2 \mid N$, and write $N = p^k M$ for some $k \geqslant 2$ and $p \nmid N$. Again using the Chinese Remainder Theorem, choose a $b \in \mathbb{Z}$ with $b \equiv 1 + p \pmod{p^k}$ and $b \equiv 1 \pmod{M}$. Then we again claim that this b is our witness, since working modulo p^2 yields:

$$b^{(N-1)/2} \equiv (1+p)^{(N-1)/2} \equiv 1 + \frac{N-1}{2} \cdot p + K \cdot p^2 \equiv 1 + \frac{N-1}{2} \cdot p \not\equiv \pm 1 \equiv \left(\frac{b}{N}\right) \pmod{p^2}.$$

This means that N cannot be an Euler pseudoprime to the base b, completing the proof. □

Corollary: Since there is some such witness, N must be an Euler pseudoprime to at least half the bases in $(\mathbb{Z}/N\mathbb{Z})^\times$. In fact, there are strictly fewer than N elements in this group, and so in fact at least half of the numbers in $\{1 \dots N\}$ are either not coprime to N (and therefore witnesses to N being composite) or are witnesses to N being composite by this criterion.

This allows us to construct an actually effective primality test!

Definition 6.7 (Solovay-Strassen Primality Test)

Given an odd integer $N > 1$ which may or may not be prime as input, the *Solovay-Strassen primality test* runs the following procedure:

1. Choose $1 < b < N$ at random.
2. Compute $d = (b, N)$ using Euclid's Algorithm (1.4). If $d > 1$, then obviously $d \mid N$, and so N is not prime. Return `composite` as output. Otherwise, progress to Step 3.
3. Compute $b^{(N-1)/2}$ modulo N and the Jacobi symbol of b on N, and check whether they are equal. If not, then N cannot be prime: again, return `composite` as output.
4. If they are congruent, then N still might be composite, but you have obtained some weak evidence that this is not true (as you would probably have returned `composite` at one of the earlier stages). Return to Step 1, choosing another b at random.

Of course, this is really a *compositionality* test, rather than a *primality* test. But if N really is composite, the probability of reaching Step 4 for the k^{th} time is strictly less than $1/2^k$.

We can set some number of rounds k as our stopping point, and eventually after that number of independent tests, return `probably prime` as output.

Remark 6.8 (Bayesian Evidence)

It is true to say that if N really is composite, the probability of being fooled by the test (seeing a false positive) k times in a row is strictly less than $1/2^k$.

Importantly, it is *not* true to say that if k rounds of this test pass, then the probability of N being composite is less than $1/2^k$.

In Bayesian terms, the prior probability of N being prime starts low. By the Prime Number Theorem (4.1), we can quantify this prior:

$$\mathbb{P}[N \text{ is prime}] \text{ can be thought of as } \pi'(N) \text{ where } \pi(N) \sim N/\log N.$$
$$= \frac{d}{dx}\frac{x}{\log x} = \frac{\log x - 1}{(\log x)^2}\bigg|_N \approx \frac{1}{\log N}.$$

Then, we double this to correct for the fact that N is odd, and all even numbers apart from 2 are composite. This gives us a prior of $\mathbb{P}[N \text{ is prime}] \approx 2/\log N$.

Every time we run the Solovay-Strassen test and do not find a witness, we obtain at least one bit of evidence that N is prime. Thus our posterior probability that N is prime is around:

$$\mathbb{P}[N \text{ is prime} \mid \text{passes } k \text{ tests}] = 1 - \frac{1 - \frac{2}{\log N}}{1 - \frac{2}{\log N} + \frac{2^{k+1}}{\log N}} = \frac{2^{k+1}}{\log N - 2 + 2^{k+1}}$$

To believe that N is prime with $1-\varepsilon$ probability requires $k_\varepsilon(N)$ tests: a function which grows approximately as $\log\log(N)$ asympotically: this is extremely slow!

In fact, we can be 99.9% sure that a randomly chosen number around 10^{200} is prime after running only 18 rounds of the Solovay-Strassen test!

How do we actually run this test? More importantly, is it even possible to compute both sides of this congruence easily? We will use the technique of *repeated squaring* to compute $b^m \pmod N$:

1. Write $m = \sum_{i=0}^{\ell} m_i 2^i$, where $m_i \in \{0, 1\}$ are the binary digits of m, and $\ell \leqslant \lceil \log_2 N \rceil$.
2. Compute b, b^2, $b^4 = (b^2)^2$, $b^8 = (b^4)^2$, and so on by squaring for ℓ steps to obtain b^{2^ℓ}.

Given these steps, we can write:

$$b^m = \prod_{i=0}^{\ell} B_i \text{ where } B_i = (b^{2^i})^{m_i} = \begin{cases} b^{2^i} & m_i = 1 \\ 1 & \text{otherwise} \end{cases}$$

which is very easy to compute. Thus we only require 2ℓ multiplications: ℓ to compute the squares, and another ℓ to compute the product given them. This is logarithmic, rather than linear, in m.

Of course, the Jacobi symbol can be computed using quadratic reciprocity (Theorem 2.16).

Definition 6.9 (Strong Pseudoprime)

Let $N \in \mathbb{N}$ be an odd composite number, and let $b \in \mathbb{Z}$ be coprime to N. Factor $N - 1 = 2^s t$, where $s \geqslant 1$ (since $N - 1$ is even) and t is odd.

We say that N is a *strong pseudoprime* to the base b if either $b^t \equiv 1$ or $b^{2^r t} \equiv -1 \pmod N$ for some $0 \leqslant r < s$.

It is not immediately obvious what the motivation for this definition is. For this, we must consider the equation $x^2 \equiv 1 \pmod p$, which has precisely two solutions ± 1 if p is an odd prime. Thus if $(a, p) = 1$ and $a^{p-1} \equiv 1 \pmod p$, then in fact $a^{(p-1)/2} \equiv \pm 1 \pmod p$.

If $(p-1)/2$ is even and $a^{(p-1)/2} \equiv 1 \pmod{p}$, then $a^{(p-1)/4} \equiv \pm 1 \pmod{p}$. We continue on, at each stage halving our exponent and checking if we have $+1$ or -1.

What happens now? At some stage, we either get -1, and then we have to stop, or we get all the way down to an odd number t (which we cannot halve) with $a^t \equiv 1 \pmod{p}$.

The notion of being a strong pseudoprime is therefore the property of behaving in the same way as a prime when we apply this repeated "square root" operation.

Proposition 6.10 (Strong Pseudoprime Properties)

If N is a strong pseudoprime to the base b, then:

(a) N is also an Euler (and hence Fermat) pseudoprime to the base b.

(b) As usual, this only depends on the value of b modulo N.

(c) If B is the set of bases b in $(\mathbb{Z}/N\mathbb{Z})^\times$ to which N is a strong pseudoprime, then the size of B is now at most a *quarter* of the size of $(\mathbb{Z}/N\mathbb{Z})^\times$.

(d) Unlike Fermat and Euler pseudoprimes, here B is *not* in general a subgroup of $(\mathbb{Z}/N\mathbb{Z})^\times$.

Proof: Omitted. □

We can use this to formulate a slightly better probabilistic primality test.

Definition 6.11 (Miller-Rabin Primality Test)

Given an odd composite integer N, the *Miller-Rabin primality test* is the algorithm:

1. Choose $1 < b < N$ at random.
2. Compute $d = (b, N)$ using Euclid's Algorithm (1.4). If $d > 1$, then obviously $d \mid N$, and so N is not prime. Return `composite` as output. Otherwise, progress to Step 3.
3. Find integers s and t such that $N - 1 = 2^s \cdot t$ and t is odd.
4. Find $c \equiv b^t \pmod{N}$. Check whether $c \equiv 1 \pmod{N}$, or if $c^{2^r} \equiv -1 \pmod{N}$ for any $0 \leqslant r < s$. If not, then N is certainly not prime: return `composite` as output.

This time, the evidence from a test which passes is twice as strong: if N is composite, there is only at most a $1/4$ chance it passes each random test!

Theorem 6.12 (Deterministic Polynomial-Time Primality Test)

Assuming that the Generalised Riemann Hypothesis. Then for any odd composite integer N, there exists a base $1 < b < 2(\log N)^2$ such that N is *not* a strong pseudoprime to the base b.

That is, the set $(\mathbb{Z}/N\mathbb{Z})^\times \setminus B$ contains at least some element less than $2(\log N)^2$.

Proof: Omitted. □

Corollary: If this is true, there is a *deterministic* polynomial-time primality test, which involves running the Miller-Rabin primality test on all the numbers up to $2(\log N)^2$: if there is no witness found among these numbers, then there is no witness anywhere, and so N must be prime!

Note: In fact, this is not even the best we can do. The Agrawal-Kayal-Saxena primality test from 2002 is unconditionally deterministic and runs in polynomial time, but is very hard to implement. Discussion of this method is well beyond the scope of this course.

Note: Polynomial-time algorithms are not necessarily faster than exponential time algorithms, only *asymptotically* faster. An exponential-time algorithm may be faster even up to a googolplex!

6.2 Fast Factorisation

Suppose that $N = ab$ is an odd composite number which is not a square. Without loss of generality, we may take $a > b > 1$, in which case we may write N as:

$$N = \left(\frac{a+b}{2}\right)^2 - \left(\frac{a-b}{2}\right)^2 = \frac{1}{4} \times \left(\left(a^2 + 2ab + b^2\right) - \left(a^2 - 2ab + b^2\right)\right) = \frac{1}{4} \times (4ab) = ab.$$

Conversely, if $N = r^2 - s^2$, then in fact $N = (r+s)(r-s)$ is a factorisation of N. This observation allows us to develop a technique for factorisation.

Definition 6.13 (Fermat Factorisation)

Suppose that N is an odd composite number which is not a square.

For each $r = \lfloor \sqrt{N} \rfloor + 1$, $r = \lfloor \sqrt{N} \rfloor + 2$, and so on, test $r^2 - N$ to see if this is a perfect square. If indeed this is equal to s^2, then $N = (r+s)(r-s)$.

Note: In fact, if this works then this will be a non-trivial factorisation of N.

Example 6.14 (Factorising 200819)

Since $\sqrt{200819} \approx 448.128$, we start with $r = 448 + 1$.

- $449^2 - 200819 = 201601 - 200819 = 782$, which is not a square.
- $450^2 - 200819 = 202500 - 200819 = 1681 = 41^2$, so take $r = 450$ and $s = 41$.

Indeed, $(r+s)(r-s) = (450+41)(450-41) = 491 \times 409 = 200819$.

Note: Fermat factorisation yields a factorisation of $N = ab$ in at most $\frac{1}{2}(a-b)$ steps. However, unfortunately in general this is no better than trying random divisors asymptotically.

Proposition 6.15 (Congruent Squares Rule)

Observe that if $N = r^2 - s^2$, then in fact $r^2 \equiv s^2 \pmod{N}$.

Suppose that $x^2 \equiv y^2 \pmod{N}$, but $x \not\equiv \pm y \pmod{N}$. Then $(N, x+y)$ and $(N, x-y)$ are both non-trivial factors of N.

Proof: Firstly, $(N, x-y)$ is a factor of N by definition. If $(N, x-y) = N$, then $x \equiv y \pmod{N}$, which we stipulated was not the case. If $(N, x-y) = 1$, then $(x+y)(x-y) \equiv 0 \pmod{N}$, and so $x + y \equiv 0 \pmod{N}$, thus $x \equiv -y \pmod{N}$, which is also false.

The same argument holds for $x + y$, which completes the proof. □

How do we turn this into a way to factorise N? We want to choose many integers x_i such that $x_i^2 \equiv c_i \pmod{N}$, where c_i has only small prime factors. Then, we can choose some subset of the x_i such that the product of the corresponding c_i is a square, and hope that the hypothesis of the above proposition holds so that we can find non-trivial factors for N.

Proposition 6.16 (Easy c_i Finding)

Suppose that $\{p_1, \ldots, p_r\}$ are primes and that $\{c_1, \ldots, c_k\}$ is a set of non-zero integers, all of whose prime factors lie exclusively in the set of primes.

Then, if $k > r + 1$, there exists some non-empty subset $I \subseteq \{1, \ldots, k\}$ such that $c_I = \prod_{i \in I} c_i$ is a square number.

Proof: We can write $c_I = m^2 \cdot \prod_{j \in S} p_j$, where I is such a subset and $S_I \subseteq S = \{0, \ldots, r\}$. Here, we use the convention that $p_0 = -1$.

There are 2^k possible subsets I, and at most 2^{r+1} possibilities for S_I. Since $k > r+1$, we know by the pigeonhole principle that there exists some pair $I \neq I'$ with $S_I = S_{I'}$. Then $c_I c_{I'}$ is a square, and we can write this as $c_{(I \triangle I')} \cdot c_{(I \cap I')} \cdot c_{(I \cap I')}$. But then $c_{(I \triangle I')}$ is a square! □

Definition 6.17 (Factor Base)

A *factor base* is a set $B \in \{-1, p_1, \ldots, p_r\}$ of -1 along with r primes. Fix an odd composite integer N. Then a *B-number* is an integer $x \in \mathbb{N}$ such that all the prime factors of $\langle x^2 \rangle$ are contained in the factor base B.

Here $\langle x^2 \rangle \equiv x^2 \pmod N$ and $-N/2 \leqslant \langle x^2 \rangle < N/2$, as in Definition 2.5.

Definition 6.18 (Factor Base Factorisation)

Let N be an odd composite integer. Choose a factor base B, and generate some B-numbers $x_1, \ldots, x_k$. Find a non-empty subset $I \subseteq \{1, \ldots, k\}$ such that the product of the $\langle x_i^2 \rangle = y^2$ is a square. (This is not obvious: just because x^2 is a square does not mean $\langle x_i^2 \rangle$ is.)

Now, let x be the product of the x_i. Then $x^2 \equiv y^2 \pmod N$. If $x \not\equiv \pm y \pmod N$, then we can find a non-trivial factor of N. Otherwise, go back and choose different B-numbers.

Note: Heuristically, if N has t distinct prime factors, then $x^2 \equiv 1 \pmod N$ can be split up into t congruences, and thus there are 2^t solutions modulo N. Now, x/y is a random solution to this congruence, and we "win" unless it is ± 1, which happens with probability $1/2^{t-1}$. This is quite good: if $t \geqslant 2$, then we should find a factorisation quickly. Thankfully, it is easy to check that N is not a perfect power p^k in polynomial time.

How do we generate B-numbers? We consider $x = \lfloor \sqrt{kN} \rfloor$ and $x = \lfloor \sqrt{kN} \rfloor + 1$ for $k \in \mathbb{N}$. This is because x^2 should be close to kN, since $\langle x^2 \rangle$ will be close to zero.

Example 6.19 (Factorising 1829)

Suppose we want to factorise 1829. Our factor base will be $B = \{-1, 2, 3, 5, 7, 11, 13\}$.

We have $\lfloor \sqrt{1829k} \rfloor = 42$, 60, 74, and 85 for $1 \leqslant k \leqslant 4$.

$\langle 42^2 \rangle = -65 = -1 \times 5 \times 13$ is a B-number. ✓

$\langle 43^2 \rangle = +20 = 2^2 \times 5$ is a B-number. ✓

$\langle 60^2 \rangle = -58 = -1 \times 2 \times 29$ is not a B-number. ×

$\langle 61^2 \rangle = +63 = 3^2 \times 7$ is a B-number. ✓

$\langle 74^2 \rangle = -11 = -1 \times 11$ is a B-number. ✓

$\langle 75^2 \rangle = +138 = 2 \times 3 \times 23$ is not a B-number. ×

$\langle 85^2 \rangle = -91 = -1 \times 7 \times 13$ is a B-number. ✓

$\langle 86^2 \rangle = +80 = 2^4 \times 5$ is a B-number. ✓

By inspection, $\langle 42^2 \rangle \times \langle 43^2 \rangle \times \langle 61^2 \rangle \times \langle 85^2 \rangle = (-5 \times 13)(2^2 \times 5)(3^2 \times 7)(-7 \times 13)$, which can be written as $2^2 \times 3^2 \times 5^2 \times 7^2 \times 13^2 = 2730^2$. But then $42 \times 43 \times 61 \times 85 \equiv 1459 \pmod N$ and $2 \times 3 \times 5 \times 7 \times 13 \equiv 901$, with $1459^2 \equiv 1554 = 901^2 \pmod N$.

Indeed, $(1829, 1459 + 901) = 59$ and $(1829, 1459 - 901) = 31$ are non-trivial factors of 1829.

Unfortunately, this isn't the whole story. To decide whether $\langle x^2 \rangle$ is a B-number, we needed to factorise it, which seems circular. However, thankfully this is easy: we need only try dividing by the elements of B. This is fairly fast, so does not pose a problem.

Also, we showed that if $k > r+1$, then a valid choice of I must exist, using the pigeonhole principle. Obviously, this is not a constructive proof: in practice, we find these with linear algebra over $\mathbb{Z}/2\mathbb{Z}$.

$$\langle x_i^2 \rangle = m^2 \prod_{j=0}^{r} p_j^{\alpha_{i,j}} \text{ where } \alpha_{i,j} \in \{0,1\}\,,\ p_0 = -1.$$

Finding I is then equivalent to finding some k-vector $\lambda \in (\mathbb{Z}/2\mathbb{Z})^k$ such that $\lambda \cdot \alpha = 0$ in this field.

Another way to generate B-numbers is by using continued fractions!

Proposition 6.20 (Convergent B-Numbers)

Let N be an odd non-square composite integer, and let p_n/q_n be a convergent of $\sqrt{N}$. Then $\left|p_n^2 - Nq_n^2\right| < 2\sqrt{N}$.

Proof: We can write this using the difference of two squares as

$$\left|p_n/q_n - \sqrt{N}\right| \left|p_n/q_n + \sqrt{N}\right| q_n^2 \leqslant \frac{q_n^2}{q_n q_{n+1}} \left(2\sqrt{N} + \frac{1}{q_n q_{n+1}}\right) = \frac{1}{q_{n+1}} \left(2q_n\sqrt{N} + \frac{1}{q_{n+1}}\right),$$

where the inequality comes from Theorem 5.7 and the triangle inequality. But since $q_n \leqslant q_{n+1} - 1$, this is in fact at most:

$$\frac{1}{q_{n+1}} \left(2q_n\sqrt{N} + \frac{1}{q_{n+1}}\right) \leqslant \frac{q_n}{q_{n+1}} \left(2\sqrt{N}\right) + \frac{1}{q_{n+1}^2} \leqslant 2\sqrt{N}$$

exactly as required. □

Corollary: Suppose now that $2\sqrt{N} < N/2$, which is equivalent to the condition $N > 16$. Then in fact $\left|p_n^2 - Nq_n^2\right| < N/2$, and so $\langle p_n^2 \rangle = p_n^2 - Nq_n^2$. Thus, since $\langle p_n^2 \rangle$ is small, it has a good chance of being a B-number! Also, since we need only compute p_n modulo N, which we can do using the recurrence $p_{n+1} = a_{n+1}p_n + p_{n-1}$ (treated as a congruence modulo N).

Example 6.21 (Factorising 12403)

The continued fraction of $\sqrt{N} = \sqrt{12403} = [111, 2, 1, 2, 2, 7, 1, \dots]$. Take the factor base to be the set $B = \{-1, 3, 13\}$. Then:

- $p_1 = 111$, so $\langle p_1^2 \rangle = -82 = -1 \times 2 \times 41$. This is not a B-number. ×
- $p_2 = 223$, so $\langle p_2^2 \rangle = 117 = 3^2 \times 13$. This is a B-number. ✓
- $p_3 = 334$, so $\langle p_3^2 \rangle = -71 = -1 \times 71$. This is not a B-number. ×
- $p_4 = 891$, so $\langle p_4^2 \rangle = 89 = 89$. This is not a B-number. ×
- $p_5 = 2116$, so $\langle p_5^2 \rangle = -27 = -1 \times 3^3$. This is a B-number. ✓
- $p_6 = 3300$, so $\langle p_6^2 \rangle = 166 = 2 \times 83$. This is not a B-number. ×
- $p_7 = 5416$, so $\langle p_7^2 \rangle = -39 = -1 \times 3 \times 13$. This is a B-number. ✓

Now, we see that $\langle 223^2 \rangle \times \langle 2116^2 \rangle \times \langle 5416^2 \rangle = (3^3 \times 13)^2$. Thus if $x \equiv 223 \times 2115 \times 5416$ and $y \equiv 3^3 \times 13$, we get $11341^2 \equiv 11574 \equiv 351^2 \pmod{N}$.

As desired, we get the non-trivial factors $(N, x + y) = 79$ and $(N, x - y) = 157$.

Note: This method using continued fractions was used in 1970 to factor the 7^{th} *Fermat number* F_7, which is equal to $2^{128} + 1 \approx 3.40 \times 10^{38}$, where $128 = 2^7$. The first four of these are prime, which led Fermat to conjecture that this was true for all of them, but Euler and Clausen factored F_5 and F_6 in 1732 and 1855 respectively.

The current best techniques include the quadratic sieve and number field sieve, developed around the 1990s. However, factorisation is still generally a hard problem!

In some cases, it is possible to find special prime factors of N easily. For example, some factors of F_7 were known before 1970, just not a full factorisation.

Remark 6.22 (Pollard's $p-1$ Method)

Suppose $N = pN_0$ is an odd composite integer, where p is prime and $p \nmid N_0$. If $(a, N) = 1$, then $a^{p-1} \equiv 1 \pmod{p}$: that is, $p \mid a^{p-1} - 1$.

However, there is no reason that we necessarily have $a^{p-1} \equiv 1 \pmod{N_0}$. This means we can compute $(a^{p-1} - 1, N)$, and hope it is a non-trivial factor of N. But this seems circular: we don't know p at the start, since the whole problem is that of factorising N.

We use *Pollard's $p-1$ method*. The algorithm is as follows:

1. Fix some $m \geqslant 2$, and compute $k = \operatorname{lcm}(1, 2, \ldots, m)$.
2. Choose some $1 < a < N$ at random, and compute $d = (a, N)$. If we get lucky, $d > 1$ is a non-trivial factor of N. Otherwise, $d = 1$.
3. Compute $a^k - 1 \pmod{N}$ quickly using repeated squaring.
4. Compute $(N, a^k - 1)$, and hope that it is a non-trivial factor of N.

Why would this work? Suppose that $p \mid N$ and that $p - 1$ is divisible only by small primes. This is possible even if p is very large. Suppose in particular that any prime power dividing $p - 1$ is at most m. Then $p - 1 \mid k$, and so $a^k \equiv 1 \pmod{p}$. In particular, $p \mid (N, a^k - 1)$, and so this will be a non-trivial factor of N.

Example 6.23 (Factorising 540143)

We take $m = 8$, and compute $k = \operatorname{lcm}(1, 2, \ldots, 8) = 840$. Choose $a = 2$, which is coprime to $N = 540143$. Then $2^k \equiv 2^{840} \equiv 2^{(64+32+8+1)\times 8} \equiv 53046 \pmod{N}$.

We compute $(540143, 53046) = 421$ using Euclid's Algorithm, and indeed this is a non-trivial factor of $N = 540143$. We see that this works because $421 - 1 = 2^2 \times 3 \times 5 \times 7$, which is the product of prime powers which are all at most 8.

Note: The factorisation methods discussed here (Fermat, Factor Bases, and Pollard) are currently the best known methods available. They run in sub-exponential time, but not in polynomial time. There is a known polynomial-time factorisation algorithm, called Shor's algorithm, but it requires a quantum computer to run. As of 2024, the largest number factorised in this way was $21 = 7 \times 3$.