NUMBER THEORY, 2025

TREVOR D. WOOLEY

1. Introduction

Number theory concerns itself with studying the multiplicative and additive structure of the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ (defined via the Peano axioms or some such: this will not concern us). Frequently, number theoretic questions are better asked in the set of all integers $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \dots\}$, and better answered by making use of the rational numbers $\mathbb{Q} = \{p/q : p \in \mathbb{Z}, q \in \mathbb{N}\}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$, where more structure may become apparent.

Some form of number theory was developed by the ancient Babylonians, Egyptians and Greeks, and many modern open problems are motivated by this work.

Problem 1.1 (Egyptian Fractions)**.** *Is it true that for all integers n with $n \geqslant 2$, there exist $x, y, z \in \mathbb{N}$ satisfying the equation*

$$\frac{4}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}\ ?$$

For example, one has

$$\frac{4}{5} = \frac{1}{2} + \frac{1}{5} + \frac{1}{10}.$$

It is generally believed that the answer to this problem should be in the affirmative. The solubility of the above equation has been checked for all $n < 10^{17}$ (by E. S. Saez, 2014), and is known to hold "for almost all" natural numbers n; see R. C. Vaughan, *On a problem of Erdős, Straus and Schinzel*, Mathematika 17 (1970), 193–198.

Problem 1.2 (Riemann Hypothesis)**.** *Is it true that when x is large enough, then one has*

$$\left| \text{card}\{p \leqslant x : p \text{ is prime}\} - \int_2^x \frac{\mathrm{d}t}{\log t} \right| < x^{1/2}(\log x)^{1000000}?$$

There is a million dollar Millenial Prize from the Clay Foundation available for a proof of the conjecture that the answer here is in the affirmative. The sharpest unconditional result in this direction has the function

$$x \exp\left(-A(\log x)^{3/5}(\log\log x)^{-1/5} \right)$$

in place of $x^{1/2}(\log x)^{1000000}$, wherein A is a suitable positive constant. This was proved independently by I. M. Vinogradov and Korobov in 1958.

Problem 1.3 (Mersenne Primes)**.** *Show that there are infinitely many primes of the shape $2^p - 1$, with p a prime number.*

At the time of writing, the largest known prime is $2^{136279841}-1$, a number with 41024320 decimal digits. The primality of this number was established through the efforts of GIMPS (see Great Internet Mersenne Prime Search, at http://www.mersenne.org/) on 19th October 2024. One can check that the integer 2^n-1 can be prime only when n is prime (why?). The integers 2^p-1 with p a prime number are known as Mersenne primes, and an industry of efficient primality tests for these special numbers is reflected in the GIMPS effort. The latest discovery earned a $3000 prize, and there is $150000 for the discovery of the first 100 million digit prime number.

Problem 1.4 (ABC Conjecture). *Show that, for each $\varepsilon>0$, there exists $C_\varepsilon>0$ (depending at most on ε) such that whenever $abc\ne 0$ and $a+b+c=0$, then*

$$\max\{|a|,|b|,|c|\}\leqslant C_\varepsilon\Bigl(\prod_{p\text{ divides }abc}p\Bigr)^{1+\varepsilon}.$$

Note here that the product is taken over distinct prime divisors of a, b and c. The ABC Conjecture has many profound implications, but until very recently seemed far beyond reach. Shinichi Mochizuki has recently claimed to have proved this conjecture. However, there is considerable skepticism concerning the validity of his proof, and despite much activity attempting to verify his proof, serious problems have been identified without fixes. See

`http:\\en.wikipedia.org\wiki\Shinichi_Mochizuki`

for more.

The conjecture that assumed the label “Fermat’s Last Theorem”, famously proved by Wiles in 1995, was motivated by the work of Diophantus. Even quite modest generalisations of this conjecture remain open.

Problem 1.5 (Generalised Fermat problem). *Is it true that the equation*

$$x^n+y^n=z^n+w^n$$

has no solutions in integers x, y, z, w, n, with $n\geqslant 5$, other than the obvious ones with

$$\{x,y\}=\{\pm z,\pm w\}\quad (n\ \textit{even}),$$
$$\{x,y\}=\{z,w\}\quad \textit{or}\quad \{x+y=z+w=0\}\quad (n\ \textit{odd})?$$

It is generally believed that the answer to this problem should be in the affirmative. For some quantitative work on this problem, see T. D. Browning, *Equal sums of two kth powers*, J. Number Theory 96 (2002), 293–318.

Problem 1.6 (Goldbach Conjecture). *Is every even integer exceeding 2 a sum of two prime numbers?*

It is generally believed that the answer to this problem should be in the affirmative. It is known that “almost all” even natural numbers can indeed be written as the sum of two primes; see H. L. Montgomery and R. C. Vaughan, *The exceptional set in Goldbach’s problem*, Acta Arith. 27 (1975), 353–370 for decisive progress in the history of this problem.

A positive integer n is called *perfect* if it is equal to the sum of all of the divisors of n (itself!) smaller than n. Thus, one sees that $6=1+2+3$, 28 and 496 are all perfect.

Problem 1.7 (Odd perfect numbers)**.** *Do there exist odd perfect numbers?*

It is generally believed that the answer to this problem should be in the negative. It is known that if n is odd and perfect, then $n > 10^{1500}$, and further n has at least 101 prime factors and at least 10 distinct prime factors (see P. Ochem and M. Rao, *Odd perfect numbers are greater than* 10^{1500}, Math. Comp. 81 (2012), 1869–1877). We note that, although Wirsing showed in 1959 that for some positive number W, and all large values of x, one has

$$\text{card}\,\{n \leqslant x : n \text{ is odd and perfect}\} \leqslant x^{W/\log\log x},$$

it remains possible that there are more odd than even perfect numbers.

2. Divisibility

We begin by reviewing some basic properties of divisibility.

Definition 2.1. (i) Suppose that $a, b \in \mathbb{Z}$. We say that b **divides** a (written $b|a$) when there exists $c \in \mathbb{Z}$ such that $a = bc$. In such circumstances, we say that a is **divisible** by b, or that b is a **divisor** of a;
(ii) When a is not divisible by b, we write $b \nmid a$;
(iii) When $b|a$ and $1 \leqslant b < a$, we say that b is a **proper divisor** of a;
(iv) We write $a^k \| b$ when $a^k | b$ but $a^{k+1} \nmid b$.
It is understood that $b|a$ makes sense only when b is non-zero.

Note that the notation $a^k\|b$ relates to the ordered pair (a, k) and b. Thus the statement $4\|24$, which is implicitly asserting that $4^1\|24$, holds because $4|24$ but $4^2 \nmid 24$. Meanwhile, the (distinct) statement $2^2\|24$ is false. In fact one has $2^3\|24$ because $2^3|24$ but $2^4 \nmid 24$. This notation is mostly used regarding prime power divisibility, and so any possible confusion will be easily avoided.

The next theorem records the basic properties of divisibility that are intuitively clear, but easily established from the definition.

Theorem 2.2. *(i) $a|a$ for every $a \in \mathbb{Z} \setminus \{0\}$;*
(ii) $a|0$ for every $a \in \mathbb{Z} \setminus \{0\}$;
(iii) if $a|b$ and $b|c$, then $a|c$;
(iv) if $a|b$ and $a|c$, then for all $x, y \in \mathbb{Z}$, one has $a|(bx + cy)$;
(v) if $a|b$ and $b|a$, then $a = \pm b$;
(vi) if $a|b$ and $a > 0$ and $b > 0$, then $a \leqslant b$;
(vii) when $m \neq 0$, one has $a|b \Leftrightarrow ma|mb$.

Proof. We will leave these assertions as exercises, though in order to illustrate ideas, we will give a formal proof of part (vii). Suppose that $m \neq 0$ and $a|b$. Then there exists $c \in \mathbb{Z}$ with the property that $b = ac$, whence $mb = m(ac)$. So there exists $c \in \mathbb{Z}$ with the property that $(mb) = (ma)c$, whence by the definition of divisibility $(ma)|(mb)$. Conversely, if $m \neq 0$ and $ma|mb$, then there exists $c \in \mathbb{Z}$ with $mb = (ma)c$. But since $m \neq 0$, the latter implies that $b = ac$. So there exists $c \in \mathbb{Z}$ with the property that $b = ac$, so from the definition of divisibility, one has $a|b$. □

The next theorem underpins the development of the theory of congruences.

Theorem 2.3 (The Division Algorithm)**.** *For any $a, b \in \mathbb{Z}$ with $a > 0$, there exist unique integers q and r with $b = qa + r$ and $0 \leqslant r < a$. If, further, one has $a \nmid b$, then one has the stronger inequality $0 < r < a$.*

Proof. Let aq be the largest multiple of a not exceeding b. Then if we put $r = b - aq$, one has $r \geqslant 0$. Moreover, by hypothesis one has $a(q+1) > b$, and thus $r = b - aq < a$. This establishes the existence of the integers q and r as stated. In order to establish uniqueness, suppose that another pair q', r' satisfy analogous conditions. If $r \neq r'$, there is no loss of generality in supposing that $r < r'$. Then since $aq' + r' = b = aq + r$, one has $a(q - q') = r' - r$, whence $a|(r' - r)$ and $0 < r' - r < a$. But the latter contradicts case (vi) of Theorem 2.2 (which would imply that $r' - r \geqslant a$). Thus we find that $r = r'$, and this now leads to the equation $qa = q'a$. But a is non-zero, so $q = q'$. Thus we find that $(q, r) = (q', r')$, and this establishes uniqueness.

Finally, if $r = 0$ then $b = qa$, whence $a|b$. The final assertion of the theorem is now immediate. □

Definition 2.4. (i) Suppose that $a \in \mathbb{Z} \setminus \{0\}$ and $b, c \in \mathbb{Z}$. We say that a is a **common divisor** of b and c when $a|b$ and $a|c$;
(ii) When b and c are not both zero, the number of common divisors of b and c is finite (see Theorem 2.2(vi)), and thus we may define the **greatest common divisor** (or **highest common factor**) of b and c to be the largest common divisor. The greatest common divisor of b and c is written (b, c) (or $\gcd(b, c)$ or $\mathrm{hcf}(b, c)$);
(iii) When $g_1, \ldots, g_n$ are integers, not all zero, we similarly write $(g_1, \ldots, g_n)$ for the largest integer d satisfying the condition that $d|g_i$ $(1 \leqslant i \leqslant n)$.

We remark that it is common to refer to the integers a and b as being *coprime* when $(a, b) = 1$.

Example 2.5. One has $(0, 2) = 2$, $(1, 3) = 1$ and $(1729, 182) = 91$ (at this point one can use trial and error, observing that (a, b) must be at most $\min\{|a|, |b|\}$).

The next theorem provides a useful tool to establish simple properties of greatest common divisors.

Theorem 2.6. *If $g = (b, c)$, then there exist integers x and y with $g = bx + cy$.*

Proof. Define the integer d by setting

$$d = \min\{bu + cv \,:\, u, v \in \mathbb{Z} \text{ and } bu + cv > 0\}.$$

Also, let x and y be the values of u and v corresponding to this minimum, so that $d = bx + cy$.

We first prove that $d|b$. If to the contrary $d \nmid b$, then by the Division Algorithm (Theorem 2.3), there exist integers r and q with $b = dq + r$ and $0 < r < d$. Then

$$r = b - dq = b - q(bx + cy) = b(1 - qx) + c(-qy),$$

whence

$$r \geqslant \min\{bu + cv \,:\, u, v \in \mathbb{Z} \text{ and } bu + cv > 0\} = d.$$

This gives a contradiction, since $r < d$, and thus we find that $d|b$.

A similar argument shows that $d|c$, and thus d is indeed a common divisor of b and c, which is to say that $d \leqslant (b, c)$. But $g = (b, c)$, and so there exist integers B and C with $b = gB$ and $c = gC$. Consequently, one has $d = g(Bx + Cy)$, and hence $g|d$. Thus $g > 0$, $d > 0$ and $g|d$, so by Theorem 2.2(vi) one has $g \leqslant d$. Then one has $d \geqslant (b, c)$ in addition to the relation $d \leqslant (b, c)$ which we derived above, so that necessarily $d = (b, c)$. But then $(b, c) = bx + cy$, and this completes the proof of the theorem. □

Theorem 2.7. *The greatest common divisor of b and c is:*
(i) the least positive value of $bx + cy$, as x and y range over $\mathbb{Z}$;
(ii) the positive common divisor of b and c that is divisible by all other such divisors.

Proof. The assertion (i) is plain from Theorem 2.6. For part (ii), observe that there exist integers x and y with $(b, c) = bx + cy$. Then if $d|b$ and $d|c$, say $b = dB$ and $c = dC$, one finds that $(b, c) = d(Bx + Cy)$, whence $d|(b, c)$. So (b, c) is divisible by all other positive common divisors of b and c. □

Remark 2.8. If $g_1, \ldots, g_n$ are not all zero, then it follows as in the proof of Theorem 2.6 that there exist integers $x_1, \ldots, x_n$ with $(g_1, \ldots, g_n) = g_1x_1 + \cdots + g_nx_n$.

The criterion for determining the greatest common divisor recorded in Theorem 2.6, and (in modified form) in Theorem 2.7, provides a simple and direct approach to establishing simple properties of the greatest common divisor function.

Theorem 2.9. *Whenever $m \in \mathbb{N}$, one has $(ma, mb) = m(a, b)$.*

Proof. Making use of Theorem 2.7(i) (twice), one has

$$\begin{aligned}(ma, mb) &= \min\{max + mby : x, y \in \mathbb{Z} \text{ and } max + mby > 0\}\\ &= m \min\{ax + by : x, y \in \mathbb{Z} \text{ and } ax + by > 0\}\\ &= m(a, b).\end{aligned}$$

□

Remark 2.10. Similiarly, when $d \in \mathbb{N}$, and $d|a$ and $d|b$, one has $(a/d, b/d) = (a, b)/d$. In particular, if $g = (a, b)$, then $(a/g, b/g) = 1$.

Proof. The first assertion follows from Theorem 2.9 by means of the relation $(d(a/d), d(b/d)) = d(a/d, b/d)$, and the second is immediate from the first. □

Theorem 2.11. *Whenever a, b, m are integers with $(a, m) = (b, m) = 1$, one has $(ab, m) = 1$.*

Proof. By Theorem 2.6, there exist integers x, y, u, v with $1 = ax + my = bu + mv$. Thus we obtain

$$(ax)(bu) = (1 - my)(1 - mv) = 1 - mw,$$

say, with $w = y + v - mvy$. Consequently, one has $(ab)(xu) + mw = 1$. But then by Theorem 2.2(iv), any common divisor of ab and m divides 1. We therefore conclude that $(ab, m) = 1$. □

Theorem 2.12. *For any integer x, and for any integers a and b, not both zero, one has*

$$(a, b) = (b, a) = (a, -b) = (a, b + ax).$$

Proof. The first assertions of the theorem are plain from Theorem 2.7(i). In order to prove that $(a,b)=(a,b+ax)$, observe that by Theorem 2.6, there exist integers u and v with $(a,b)=au+bv$, whence $(a,b)=a(u-xv)+(b+ax)v$. We therefore have $(a,b+ax)|(a,b)$. But $(a,b)|a$ and $(a,b)|b$, so $(a,b)|(b+ax)$. But now we have $(a,b+ax)|(a,b)|(a,b+ax)$, and so by virtue of positivity, Theorem 2.2(v) establishes the desired conclusion. □

Example: Compute $(n^2+1,n+1)$ for $n\in\mathbb{Z}$.
Solution: Observe that repeated application of Theorem 2.12 shows that

$$\begin{aligned}(n^2+1,n+1)&=(n^2+1-n(n+1),n+1)=(1-n,n+1)\\&=(1-n+(n+1),n+1)=(2,n+1),\end{aligned}$$

whence

$$(n^2+1,n+1)=\begin{cases}2, & \text{when } n \text{ is odd,}\\ 1, & \text{when } n \text{ is even.}\end{cases}$$

Theorem 2.13. *Suppose that $c|ab$ and $(b,c)=1$. Then $c|a$.*

Proof. By Theorem 2.9, the hypotheses of the theorem imply that $(ab,ac)=|a|(b,c)=|a|$. But by hypothesis, one has $c|ab$, which implies that $c|(ab,ac)$. We thus conclude that $c|a$. □

At last we are positioned to describe an algorithm for calculating greatest common divisors. Of course, by exhaustive checking one could determine the greatest common divisor of two integers b and c in time $O(\min\{|b|,|c|\})$, but the Euclidean Algorithm has running time only $O(\log(\min\{|b|,|c|\}))$. Indeed, for most pairs of integers b and c, the Euclidean Algorithm takes only about $(12\log 2/\pi^2)\log(\max\{|b|,|c|\})$ steps.

Theorem 2.14 (Euclidean Algorithm)**.** *Suppose that $b\in\mathbb{Z}$ and $c\in\mathbb{N}$. Define the integers r_i and q_i for $i\geqslant 1$ by repeated application of the Division Algorithm thus:*

$$\begin{aligned}b&=cq_1+r_1, &&\text{with } 0<r_1<c,\\ c&=r_1q_2+r_2, &&\text{with } 0<r_2<r_1,\\ r_1&=r_2q_3+r_3, &&\text{with } 0<r_3<r_2,\\ &\dots\\ r_{j-2}&=r_{j-1}q_j+r_j, &&\text{with } 0<r_j<r_{j-1},\\ r_{j-1}&=r_jq_{j+1}.\end{aligned}$$

(Here we adopt obvious conventions if the process terminates prematurely.) Then $(b,c)=r_j$, the last non-zero remainder in the division process.

Proof. Repeated application of Theorem 2.12 yields

$$\begin{aligned}(b,c)&=(b-cq_1,c)=(r_1,c)\\&=(c-r_1q_2,r_1)=(r_2,r_1)\\&=(r_1-r_2q_3,r_2)=(r_3,r_2)\\&=\dots=(r_j,r_{j-1})=(r_j,0)=r_j.\end{aligned}$$

This conclusion of the theorem follows at once. □

Observation 2.15. *One can apply the Euclidean Algorithm to obtain integral solutions (x, y) to linear equations of the shape $bx + cy = (b, c)$ by "reversing" the application of the algorithm. In general, one can apply this method to solve the equation $bx + cy = k$ whenever $(b, c)|k$ (Why? Convince yourself that this is the case.)*

Proof. Using the notation employed in the statement of the Euclidean Algorithm, one finds that r_1 is a linear combination of b and c, and then that r_2 is a linear combination of c and r_1, and hence of b and c, and that r_3 is a linear combination of r_1 and r_2, and hence of b and c, and so on. In this way, we see that every remainder r_i that occurs in the algorithm is itself a linear combination of b and c, and the desired conclusion follows. □

Example 2.16. Determine the greatest common divisor of 2025 and 323, and find integers x and y with $2025x + 323y = (2025, 323)$.

Proof. Applying the Euclidean Algorithm, we obtain

$$\begin{aligned} 2025 &= 323 \cdot 6 + 87 \\ 323 &= 87 \cdot 3 + 62 \\ 87 &= 62 \cdot 1 + 25 \\ 62 &= 25 \cdot 2 + 12 \\ 25 &= 12 \cdot 2 + 1 \\ 12 &= 12 \cdot 1, \end{aligned}$$

and so $(2025, 323) = 1$. Reversing this application of the Euclidean Algorithm, we find that

$$\begin{aligned} 1 &= 25 - 12 \cdot 2 \\ &= 25 - (62 - 25 \cdot 2) \cdot 2 = 25 \cdot 5 - 62 \cdot 2 \\ &= (87 - 62 \cdot 1) \cdot 5 - 62 \cdot 2 = 87 \cdot 5 - 62 \cdot 7 \\ &= 87 \cdot 5 - (323 - 87 \cdot 3) \cdot 7 = 87 \cdot 26 - 323 \cdot 7 \\ &= (2025 - 323 \cdot 6) \cdot 26 - 323 \cdot 7 = 2025 \cdot 26 - 323 \cdot 163. \end{aligned}$$

Thus, the equation $2025x + 323y = (2025, 323) = 1$ has the solution $(x, y) = (26, -163)$. □

Note 2.17. *One can obtain integral solutions to linear equations in more variables by breaking the equation down into subequations of two variables each. In order to illustrate the strategy, consider the equation $18x + 39y + 77z = 1$. One can verify easily that $(18, 39) = 3$, and so the equation $18x + 39y = 3$ possesses an integral solution, say $18x_0 + 39y_0 = 3$, which may be found via the Euclidean Algorithm. Now substitute this solution into the original equation with an additional parameter, and solve the resulting equation. We obtain the equation $3l + 77z = 1$. Since $(3, 77) = 1$, the latter equation has an integral solution $(l, z) = (l_0, z_0)$, say, which may be found via the Euclidean Algorithm. A solution of the original equation is then given by $(x, y, z) = (l_0x_0, l_0y_0, z_0)$.*

We finish this section by introducing the concept of least common multiples.

Definition 2.18. (i) Non-zero integers $a_1, \ldots, a_n$ are said to have a *common multiple* b when $a_i|b$ for $1 \leqslant i \leqslant n$.

(ii) The *least common multiple* of the non-zero integers $a_1, \ldots, a_n$ is the smallest positive common multiple of these integers, which we denote by $[a_1, \ldots, a_n]$.

Theorem 2.19. *(i) If m is a positive integer and a and b are non-zero integers, then $[ma, mb] = m[a, b]$.*
(ii) When a and b are non-zero integers, one has $[a, b](a, b) = |ab|$.

Proof. First consider the assertion of part (i) of the theorem. Let $D = [ma, mb]$ and $d = [a, b]$. Then md is a multiple of both ma and mb, so that $md \geqslant D$. Also, D is a multiple of both ma and mb, so that D/m is a multiple of both a and b. Then $D/m \geqslant d$. We have therefore shown that $md \leqslant D \leqslant md$, whence $D = md$. This establishes part (i) of the theorem.

Now consider part (ii). Put $d = (a, b)$. Then $(a/d, b/d) = (a, b)/d = 1$ and $[a/d, b/d] = [a, b]/d$. We aim to show that whenever a' and b' satisfy $(a', b') = 1$, then $[a', b'](a', b') = |a'b'|$, for then we obtain $[a/d, b/d](a/d, b/d) = |ab|/d^2$, whence $([a, b]/d)((a, b)/d) = |ab|/d^2$, so that $[a, b](a, b) = |ab|$, as desired. There is no loss of generality in supposing that $a' > 0$ and $b' > 0$. We may suppose that $[a', b'] = ma'$, with $b'|ma'$. Since we now suppose that $(a', b') = 1$, it follows from Theorem 2.13 that $b'|m$, whence $b' \leqslant m$. Then $b'a' \leqslant ma'$. But $b'a' \geqslant [b', a'] = ma'$. We therefore conclude that $b'a' = [b', a']$ whenever $(b', a') = 1$. In view of our earlier remarks, the desired conclusion follows. □

Theorem 2.20. *Suppose that $b_1, \ldots, b_n$ are non-zero integers. Then, putting $k = [b_1, \ldots, b_n]$, the set of all common multiples of the integers $b_1, \ldots, b_n$ is given by $\{km : m \in \mathbb{Z}\}$.*

Proof. Exercise. □

3. Primes and the fundamental theorem of arithmetic

Definition 3.1. A natural number p satisfying the conditions (i) $p > 1$, and (ii) that whenever $d|p$, one has $|d| = 1$ or p, is called a **prime number**. Any integer exceeding 1 which is not a prime number is called a **composite number**.

Theorem 3.2 (Factorisation into primes)**.** *Every integer n exceeding 1 may be written as a product of prime numbers.*

Proof. The theorem plainly holds for $n = 2$. Suppose that the theorem holds for $1 < n \leqslant N$. The least divisor d of $N + 1$ with $d > 1$ is plainly prime, say p. But $(N + 1)/p \leqslant N$, so is either equal to 1, or else by hypothesis is a product of prime numbers. Then $N + 1$ is also a product of prime numbers. Consequently, by induction, we find that all integers exceeding 1 are a product of prime numbers. □

Given a factorisation of an integer n into prime numbers, one may collect together like primes and order the primes by size so as to give a factorisation

$$n = \pm \prod_{i=1}^{s} p_i^{r_i},$$

where $p_1 < p_2 < \cdots < p_s$ are prime numbers, and $r_i \in \mathbb{N}$ $(1 \leqslant i \leqslant s)$. We will call this the *canonical prime factorisation* of n. Note that the empty product of (no) primes is equal to 1. If the choice of sign, the primes p_i, and the exponents r_i, are uniquely determined, we say that n has a *unique factorisation* into primes.

Lemma 3.3. *Suppose that p is a prime number, and $p|a_1 \dots a_t$. Then $p|a_i$ for some i with $1 \leqslant i \leqslant t$.*

Proof. We prove first that if m and n are natural numbers and $p|mn$, then $p|m$ or $p|n$. For if $p \nmid m$, then $(p, m) = 1$, and then it follows from Theorem 2.13 that $p|n$. Moving now to the general case, the latter argument shows that when $p|a_1 \dots a_t$, then either $p|a_1$ or $p|a_2 \dots a_t$. The conclusion of the lemma therefore follows by induction on t. □

Theorem 3.4 (The Fundamental Theorem of Arithmetic)**.** *Positive integers $n > 1$ have unique factorisations into primes.*

Proof. Suppose, by way of deriving a contradiction, that $n > 1$ is the smallest natural number that fails to have a unique factorisation into primes. Let p be a prime factor of n. It follows from Lemma 3.3 that all factorisations of n contain p as one of the prime factors. One cannot have $p = n$, since then n factors uniquely into primes. Consequently, the integer $n_0 = n/p$ satisfies $1 < n_0 < n$, and hence possesses a unique factorisation into primes. But then $n = pn_0$ likewise has a unique factorisation into primes, contradicting our opening hypothesis. It therefore follows that all positive integers $n > 1$ have a unique factorisation into primes. □

Remark 3.5. The unique factorisation theorem enables one to determine greatest common divisors and least common multiples simply. At least, that is the case when prime factorisations are available, which is computationally expensive data to assemble (the Euclidean Algorithm, on the other hand, is computationally very cheap). Suppose that

$$a = \prod_{i=1}^{s} p_i^{r_i} \quad \text{and} \quad b = \prod_{i=1}^{s} p_i^{t_i},$$

with the p_i distinct prime numbers and the exponents r_i and t_i non-negative integers. Then on has

$$(a, b) = \prod_{i=1}^{s} p_i^{\min\{r_i, t_i\}} \quad \text{and} \quad [a, b] = \prod_{i=1}^{s} p_i^{\max\{r_i, t_i\}}.$$

Moreover, since $\min\{r_i, t_i\} + \max\{r_i, t_i\} = r_i + t_i$, it follows from the latter formulae that $(a, b)[a, b] = |ab|$, as has already been established in Theorem 2.19(ii).

Theorem 3.6 (Euclid)**.** *There are infinitely many prime numbers, and hence also arbitrarily large prime numbers.*

Proof. Suppose to the contrary that there are only finitely many prime numbers, say $p_1, \dots, p_n$. None of $p_1, \dots, p_n$ divides the auxiliary integer $Q_n = p_1 \dots p_n + 1$, so either Q_n is itself prime, or else it is divisible by a prime different from $p_1, \dots, p_n$. This yields a contradiction, and the theorem follows. □

Note that, writing p_n for the n-th prime number, the expression $p_1 p_2 \dots p_n + 1$ is not always prime. Thus, for example, we have $2 \cdot 3 \cdot \ldots \cdot 13 + 1 = 30031 = 59 \cdot 509$. It is conjectured that, starting with $q_1 = 2$, if one defines q_{n+1} to be the least prime divisor of the integer $q_1 \dots q_n + 1$, then the sequence (q_n) should consist of all the prime numbers. A forthcoming homework exercise proves a result related to this conjecture. Thus, writing p_n for the n-th prime, one can prove that p_{n+1} is the smallest prime divisor of the integer

$$(p_1 \dots p_n)^{(p_1 \cdots p_n)^{(p_1 \cdots p_n)}} - 1.$$

Theorem 3.7. *The n-th smallest prime number p_n satisfies $p_n < 2^{2^n}$.*

Proof. Since $p_1 = 2$, the conclusion claimed in the theorem holds for $n = 1$. Suppose that N is a natural number, and that the conclusion holds when $1 \leqslant n \leqslant N$. Then by the argument of the proof of Theorem 3.6, one finds that

$$p_{N+1} \leqslant p_1 p_2 \ldots p_N + 1 < 2^{2^1} 2^{2^2} \ldots 2^{2^N} + 1 < 2^{2^{N+1}-1} + 1 < 2^{2^{N+1}}.$$

Then the conclusion holds also for $N + 1$, and so the desired conclusion follows by induction. □

Now define the function $\pi(x)$ for positive numbers x by putting

$$\pi(x) = \sum_{\substack{p \leqslant x \\ p \text{ prime}}} 1.$$

Thus one has $\pi(2) = 1$, $\pi(3) = 2$, $\pi(\sqrt{10}) = 2$, and so on.

Corollary 3.8. *One has $\pi(x) > \log\log x$ for $x \geqslant 2$.*

Proof. One can verify this assertion using the conclusion of Theorem 3.7. □

The "exercise" at the end of this section shows that there are constants c_1 and c_2 with $0 < c_1 < 1 < c_2$ such that for each number x with $x \geqslant 2$, one has

$$c_1 x / \log x \leqslant \pi(x) \leqslant c_2 x / \log x.$$

In fact, using complex analysis and the Riemann zeta function, defined for $\mathrm{Re}(s) > 1$ by means of the series

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p (1 - p^{-s})^{-1},$$

and by analytic continuation for $s \neq 1$, one can prove that

$$\pi(x) \sim x / \log x, \quad \text{as} \quad x \to \infty.$$

This asymptotic formula was proved by Hadamard and de la Vallée Poussin in 1896 (for a detailed account of this work, see E. C. Titchmarsh, *The theory of the Riemann zeta-function.* Second edition. Edited and with a preface by D. R. Heath-Brown. The Clarendon Press, Oxford University Press, Oxford, 1986). Thus the nth prime number has size about $n \log n$. One of the list of 7 Millenial Problems proposed by the Clay Mathematics Institute is the resolution of the Riemann Hypothesis, which asserts that the analytic continuation of $\zeta(s)$ to the complex plane has, aside from the trivial zeros at $s = -2$, -4, ..., only zeros on the half-line $\mathrm{Re}(s) = 1/2$ (see http://www.claymath.org/millennium). An accessible conclusion equivalent to this assertion is that a positive number C exists for which the upper bound

$$\left| \pi(x) - \int_2^x \frac{\mathrm{d}t}{\log t} \right| < C x^{1/2} (\log x)^{1000000}$$

holds for $x > 2$. The sharpest unconditional result in this direction has the function

$$x \exp\left(-A(\log x)^{3/5} (\log\log x)^{-1/5}\right)$$

in place of $x^{1/2}(\log x)^{1000000}$, wherein A is a suitable positive constant. This was proved independently by I. M. Vinogradov and Korobov in 1958 (see the book by Titchmarsh for an account of this work).

Given an interesting sequence such as the prime numbers, number theorists are interested in analysing features of their distribution. We begin with arithmetic progressions, about which we will say more as the course progresses.

Theorem 3.9. *There are infinitely many prime numbers of the shape $4k+3$, with k a non-negative integer.*

Proof. Suppose that there are only finitely many prime numbers of the shape $4k+3$ with $k \geqslant 0$, say $p_1, \ldots, p_n$. Consider the integer $Q = 4p_1 \ldots p_n - 1$. The integer Q is odd, and of the shape $4k+3$, so cannot be divisible exclusively by primes of the shape $4k+1$. Moreover, none of the primes $p_1, \ldots, p_n$ divide Q. Thus Q is divisible by a new prime of the shape $4k+3$ not amongst $p_1, \ldots, p_n$, contradicting our initial hypothesis. This completes the proof of the theorem. □

One can imitate the above proof to show that there are infinitely many prime numbers of the shape $6k+5$ ($k \in \mathbb{N}$). But the corresponding proof for $4k+1$ is not so easy. What about proving that there are infinitely many primes of the shape $5k+4$? See a forthcoming homework problem for a proof that there are infinitely many prime numbers of the shape $4k+1$ ($k \in \mathbb{N}$).

At the time of writing, the largest known prime is $2^{136279841} - 1$, a number with 41024320 decimal digits. The primality of this number was established through the efforts of GIMPS (see Great Internet Mersenne Prime Search, at http://www.mersenne.org/) on 19th October 2024. One can check that the integer $2^n - 1$ can be prime only when n is prime (why?). The integers $2^p - 1$ with p a prime number are known as Mersenne primes, and an industry of efficient primality tests for these special numbers is reflected in the GIMPS effort. On the other hand, it is conjectured that there are only finitely many Fermat primes, that is to say, integers of the shape $2^{2^n} + 1$ which are prime numbers. These integers are known to be prime for $n = 0, 1, 2, 3, 4$, and at the time of writing known to be composite for $5 \leqslant n \leqslant 32$.

Now we consider gaps between consecutive prime numbers.

Theorem 3.10. *There are arbitrarily large gaps between consecutive prime numbers.*

Proof. Consider the sequence $n! + 2$, $n! + 3$, ..., $n! + n$ of $n - 1$ consecutive integers. The first of these integers is divisible by 2, the second by 3, and so on, with the last divisible by n. None of these integers can be prime, therefore, and so there are gaps of length at least $n - 1$, for any natural number n, between consecutive prime numbers. □

This theorem shows that one can find gaps between consective primes p_n and p_{n+1} at least as large as $C \log p_n / \log\log p_n$, for a suitable positive constant C, infinitely often. It was shown in December 2014 by Ford, Green, Konyagin, Maynard and Tao that there is a positive number C with the property that the gaps can be as large as

$$C \frac{(\log p_n)(\log\log p_n)(\log\log\log\log p_n)}{\log\log\log p_n}$$

infinitely often (see arXiv:1412.5029). On the other hand, in July 2014 the mathematics consortium D. H. J. Polymath, led by Terry Tao, has built on the pivotal work of Yitang Zhang and James Maynard to prove that $p_{n+1} - p_n \leqslant 246$ infinitely often (see arXiv:1407.4897), thus providing an approximation to the Twin Prime Conjecture that $p_{n+1} - p_n = 2$ infinitely often.

A natural question is whether there are simple ways to produce prime numbers. The next theorem shows that polynomials, at least, cannot take prime values all the time.

Theorem 3.11. *There is no non-constant polynomial which takes only prime values.*

Proof. Suppose that $f(t)$ is a polynomial with integral coefficients. Then for every pair of large integers n and m, an examination of the Taylor expansion (which for polynomials is equivalent to a binomial expansion) reveals that $f(n)$ is a proper divisor of $f(n + mf(n))$ exceeding 1. Thus we see that $f(n + mf(n))$ is composite. □

Matijasevich showed in 1970 that there exist polynomials $f(n_1, \dots, n_k)$, all of whose positive values are prime numbers, and indeed such polynomials exist in 12 variables. Moreover, there are infinitely many prime numbers of the shape $x^2 + y^4$ (Friedlander and Iwaniec, 1998; see J. B. Friedlander and H. Iwaniec, *The polynomial $X^2 + Y^4$ captures its primes.* Ann. of Math. (2) 148 (1998), 945–1040), and also of the form $x^3 + 2y^3$ (Heath-Brown, 2001; see D. R. Heath-Brown, *Primes represented by $x^3 + 2y^3$.* Acta Math. 186 (2001), 1–84). For linear polynomials, much more is known, as we shall see later in the course. When a and b are natural numbers with $(a, b) = 1$, Dirichlet proved that $an + b$ is prime for infinitely many integers n (this was proved in 1830).

Adventurous students may wish to follow the steps below to obtain information about the asymptotic behaviour of $\pi(x)$ previously advertised:

(a) For each $n \geqslant 1$, and each prime p, prove that $p^h \| n!$, where $h = \sum_{m=1}^{\infty} \lfloor n/p^m \rfloor$, and $\lfloor z \rfloor$ denotes $\max\{n \in \mathbb{Z} : n \leqslant z\}$.

(b) Prove that, for each $x \in \mathbb{R}$, we have $\lfloor x \rfloor - 2\lfloor x/2 \rfloor \leqslant 1$. Hence prove that

$$\prod_{n<p\leqslant 2n} p \quad \text{divides} \quad (2n)!/(n!)^2 \quad \text{divides} \quad \prod_{p\leqslant 2n} p^{r_p} \quad (n \geqslant 2),$$

where r_p is the largest integer such that $p^{r_p} \leqslant 2n$. Deduce that

$$(\pi(2n) - \pi(n)) \log n \leqslant \log\left((2n)!/(n!)^2\right) \leqslant \pi(2n)\log(2n).$$

(c) Prove that $2^n \leqslant (2n)!/(n!)^2 \leqslant 2^{2n}$ for $n \geqslant 2$. Deduce that there are constants $c_1, c_2 > 0$ such that $\pi(n) > c_1 n/\log n$ and $\pi(2n) - \pi(n) < c_2 n/\log n$.

(d) Deduce from part (c) that when $y \geqslant 2$, there is a constant $c_3 > 0$ such that $\pi(y) - \pi(y/2) < c_3 y/\log y$. Infer that there is a constant $c_4 > 0$ such that $\pi(y)\log y - \pi(y/2)\log(y/2) < c_4 y$.

(e) Apply the last inequality with $y = x/2^m$ to show that when $m \geqslant 0$ and $2^m \leqslant x/2$, one has $\pi(x) < c_5 x/\log x$ for a constant $c_5 > 0$. Infer from part (c) that for $x \geqslant 2$, one has $\pi(x) > c_6 x/\log x$ for a suitable constant $c_6 > 0$. Hence there are constants $c_5 > 0$ and $c_6 > 0$ for which $c_6 x/\log x < \pi(x) < c_5 x/\log x$.

4. Congruences

We begin by introducing some definitions and elementary properties.

Definition 4.1. Suppose that $a, b \in \mathbb{Z}$ and $m \in \mathbb{N}$. We say that a is **congruent** to b modulo m, and write $a \equiv b \pmod{m}$, when $m|(a-b)$.
We say that a is **not congruent** to b modulo m, and write $a \not\equiv b \pmod{m}$, when $m \nmid (a-b)$.

Theorem 4.2. *Let a, b, c, d be integers. Then*
(i) $a \equiv b \pmod{m} \iff b \equiv a \pmod{m} \iff a - b \equiv 0 \pmod{m}$*;*
(ii) $a \equiv b \pmod{m}$ *and* $b \equiv c \pmod{m} \Rightarrow a \equiv c \pmod{m}$*;*
(iii) $a \equiv b \pmod{m}$ *and* $c \equiv d \pmod{m} \Rightarrow a + c \equiv b + d \pmod{m}$ *and* $ac \equiv bd \pmod{m}$*;*
(iv) If $a \equiv b \pmod{m}$ *and* $d|m$ *with* $d > 0$*, then* $a \equiv b \pmod{d}$*;*
(v) If $a \equiv b \pmod{m}$ *and* $c > 0$*, then* $ac \equiv bc \pmod{mc}$*.*

Proof. Try this as an exercise. You can check that congruence modulo m is an equivalence relation on $\mathbb{Z}$, and the ring properties of $\mathbb{Z}$ are preserved under congruence modulo m. □

Corollary 4.3. *When $p(t)$ is a polynomial with integral coefficients, it follows that whenever* $a \equiv b \pmod{m}$*, then* $p(a) \equiv p(b) \pmod{m}$*.*

Proof. Use induction to establish that whenever $a \equiv b \pmod{m}$, then $a^i \equiv b^i \pmod{m}$ for each $i \in \mathbb{N}$. □

The next theorem indicates how factors may be cancelled through congruences.

Theorem 4.4. *Let* $a, x, y \in \mathbb{Z}$ *and* $m \in \mathbb{N}$*. Then*
(i) $ax \equiv ay \pmod{m} \iff x \equiv y \pmod{m/(a,m)}$*;*
(ii) If $ax \equiv ay \pmod{m}$ *and* $(a,m) = 1$*, then* $x \equiv y \pmod{m}$*;*
(iii) $x \equiv y \pmod{m_i}$ $(1 \leqslant i \leqslant r) \iff x \equiv y \pmod{[m_1, \ldots, m_r]}$*.*

Proof. Observe first that when $(a,m) = 1$, then $m|a(x-y) \iff m|(x-y)$. Then the conclusion of part (ii) follows, and this also delivers part (i) whenever $(a,m) = 1$. When $(a,m) > 1$, on the other hand, one does at least have $(a/(a,m), m/(a,m)) = 1$, so that

$$m|a(x-y) \iff \frac{m}{(a,m)} \biggm| \frac{a}{(a,m)}(x-y) \iff \frac{m}{(a,m)} \biggm| (x-y).$$

This establishes the conclusion of part (i) of the theorem.

We now consider part (iii) of the theorem. Observe first that whenever $m_i|(x-y)$ for $(1 \leqslant i \leqslant r)$, then $[m_1, \ldots, m_r]|(x-y)$. On the other hand, if $[m_1, \ldots, m_r]|(x-y)$, then $m_i|(x-y)$ for $(1 \leqslant i \leqslant r)$. The conclusion of part (iii) is now immediate. □

Now we examine the set of equivalence classes with respect to congruence modulo m.

Definition 4.5. (i) If $x \equiv y \pmod{m}$, then y is called a **residue** of x modulo m;
(ii) We say that $\{x_1, \ldots, x_m\}$ is a **complete residue system** modulo m if for each $y \in \mathbb{Z}$, there exists a unique x_i with $y \equiv x_i \pmod{m}$;
(iii) The set of integers x with $x \equiv a \pmod{m}$ is called the **residue class**, or **congruence class**, of a modulo m.

We also wish to consider residue classes containing integers coprime to the modulus, and this prompts the following observation.

Theorem 4.6. *Whenever $b \equiv c \pmod{m}$, one has $(b, m) = (c, m)$.*

Proof. If $b \equiv c \pmod{m}$, then $m|(b-c)$, whence there exists an integer x with $b = c+mx$. But then $(b, m) = (c + mx, m) = (c, m)$, as desired. □

Definition 4.7. (i) A **reduced residue system** modulo m is a set of integers $r_1, \dots, r_n$ satisfying (a) $(r_i, m) = 1$ for $1 \leqslant i \leqslant n$, (b) $r_i \not\equiv r_j \pmod{m}$ for $i \neq j$, and (c) whenever $(x, m) = 1$, then $x \equiv r_i \pmod{m}$ for some i with $1 \leqslant i \leqslant n$;
(ii) The number of elements in a reduced residue system modulo m is denoted by $\phi(m)$ (Euler's totient, or Euler's ϕ-function).

Theorem 4.8. *The number $\phi(m)$ is equal to the number of integers n with $1 \leqslant n \leqslant m$ and $(n, m) = 1$.*

Proof. This is immediate from the definition of the Euler totient. □

Theorem 4.9. *Suppose that $(a, m) = 1$. Then whenever $\{r_1, \dots, r_n\}$ is a complete (respectively, reduced) residue system modulo m, the set $\{ar_1, \dots, ar_n\}$ is also a complete (respectively, reduced) residue system modulo m.*

Proof. When $(a, m) = 1$, it follows from Theorem 4.4(ii) that

$$ar_i \equiv ar_j \pmod{m} \iff r_i \equiv r_j \pmod{m}.$$

Hence the sets $\{r_1, \dots, r_n\}$ and $\{ar_1, \dots, ar_n\}$ are in bijective correspondence. Thus $\{ar_1, \dots, ar_n\}$ must be a complete residue system whenever $\{r_1, \dots, r_n\}$ is such (because these sets have the same number of elements). Moreover, since $(a, m) = 1$, it follows that whenever $(r_i, m) = 1$ one has $(ar_i, m) = 1$, and so each element ar_i is a reduced residue. Since the two sets in question have the same number of elements, we find that whenever $\{r_1, \dots, r_n\}$ is a reduced residue system, then so is $\{ar_1, \dots, ar_n\}$. □

Theorem 4.10 (Euler, 1760)**.** *If $(a, n) = 1$, then $a^{\phi(n)} \equiv 1 \pmod{n}$.*

Proof. Let $\{r_1, r_2, \dots, r_{\phi(n)}\}$ be any reduced residue system modulo n, and suppose that $(a, n) = 1$. By Theorem 4.9, the system $\{ar_1, \dots, ar_{\phi(n)}\}$ is also a reduced residue system modulo n. Then there is a permutation σ of $\{1, 2, \dots, \phi(n)\}$ with the property that $r_i \equiv ar_{\sigma i} \pmod{n}$ $(1 \leqslant i \leqslant \phi(n))$. Consequently, one has

$$\prod_{i=1}^{\phi(n)} r_i \equiv \prod_{i=1}^{\phi(n)} (ar_{\sigma i}) \equiv \prod_{j=1}^{\phi(n)} (ar_j) \equiv a^{\phi(n)} \prod_{j=1}^{\phi(n)} r_j \pmod{n}.$$

But $(r_1 \dots r_{\phi(n)}, n) = 1$, and thus $a^{\phi(n)} \equiv 1 \pmod{n}$. □

Corollary 4.11 (Fermat's Little Theorem, 1640)**.** *Let p be a prime number, and suppose that $(a, p) = 1$. Then one has $a^{p-1} \equiv 1 \pmod{p}$. Meanwhile, for all integers a one has $a^p \equiv a \pmod{p}$.*

Proof. Note that the set $\{1, 2, \dots, p - 1\}$ is a reduced residue system modulo p. Thus $\phi(p) = p-1$, and the first part of the theorem follows from Theorem 4.10. When $(a, p) = 1$, the second part of the theorem is immediate from the first part. Meanwhile, if $(a, p) > 1$, one has $p|a$, and then one plainly has $a^p \equiv a \pmod{p}$. This completes the proof of the theorem. □

Fermat's Little Theorem, and Euler's Theorem, ensure that the computation of powers is very efficient modulo p (or modulo m).

Example 4.12. Compute $5^{2025} \pmod{41}$. Observe first that $\phi(41) = 40$, and so it follows from Fermat's Little Theorem that $5^{40} \equiv 1 \pmod{41}$, and hence

$$5^{2025} = (5^{40})^{50} 5^{25} \equiv 5^{25} \pmod{41}.$$

Note next that powers which are themselves powers of 2 are easy to compute by repeated squaring (the "divide and conquer" algorithm). Thus one finds that $5^2 \equiv 25 \equiv -16 \pmod{41}$, $5^4 = (5^2)^2 \equiv (-16)^2 \equiv 10 \pmod{41}$, $5^8 \equiv (5^4)^2 \equiv (10)^2 \equiv 18 \pmod{41}$, $5^{16} \equiv (5^8)^2 \equiv 18^2 \equiv 324 \equiv -4 \pmod{41}$. In this way we deduce that

$$5^{2025} \equiv 5^{16} \cdot 5^8 \cdot 5^1 \equiv (-4) \cdot 18 \cdot 5 \equiv -360 \equiv 9 \pmod{41}.$$

The strategy of computing power of 2 powers of residues is one that is effective in general. The residue of $a^r \pmod{m}$ may be computed by writing the base 2 expansion of r, computing the relevant power of 2 powers that occur in this binary expansion by repeated squaring, and then multiplying together to obtain the rth power.

Euler's Theorem provides one (rather inefficient) method of computing multiplicative inverses modulo m. An efficient method is based on the Euclidean Algorithm.

Theorem 4.13. *Suppose that $(a, m) = 1$. Then there exists an integer x with the property that $ax \equiv 1 \pmod{m}$. If x_1 and x_2 are any two such integers, then $x_1 \equiv x_2 \pmod{m}$. Also, if $(a, m) > 1$, then there exists no integer x with $ax \equiv 1 \pmod{m}$.*

Proof. Suppose that $(a, m) = 1$. Then by the Euclidean Algorithm, there exist integers x and y such that $ax + my = 1$, whence $ax \equiv 1 \pmod{m}$. Meanwhile, if $ax_1 \equiv 1 \equiv ax_2 \pmod{m}$, then $a(x_1 - x_2) \equiv 0 \pmod{m}$. But $(a, m) = 1$, and thus $x_1 - x_2 \equiv 0 \pmod{m}$. We have therefore established both existence and uniqueness of the multiplicative inverse for residues a with $(a, m) = 1$. If $(a, m) > 1$, then $(ax, m) > 1$ for every integer x. But if one were to have $ax \equiv 1 \pmod{m}$, then $(ax, m) = (1, m) = 1$, which yields a contradiction. This establishes the last part of the theorem. □

We have just shown that the congruence classes of a reduced residue system modulo m form a group under multiplication modulo m.

Theorem 4.14 (Wilson's Theorem; Waring 1770, Lagrange)**.** *For each prime number p, one has $(p - 1)! \equiv -1 \pmod{p}$.*

Proof. The proof for $p = 2$ and 3 is immediate, so suppose henceforth that p is a prime number with $p \geqslant 5$. Observe that when $1 \leqslant a \leqslant p - 1$, one has $(a, p) = 1$, so there exists an integer $\overline{a}$ unique modulo p with $a\overline{a} \equiv 1 \pmod{p}$. Moreover, there is no loss in supposing that $\overline{a}$ satisfies $1 \leqslant \overline{a} \leqslant p-1$, and then $\overline{a}$ is a uniquely defined integer. We may now pair off the integers a with $1 \leqslant a \leqslant p-1$ with their counterparts $\overline{a}$ with $1 \leqslant \overline{a} \leqslant p-1$, so that $a\overline{a} \equiv 1 \pmod{p}$ for each pair. Note that $a \neq \overline{a}$ so long as $a^2 \not\equiv 1 \pmod{p}$. But $a^2 \equiv 1 \pmod{p}$ if and only if $(a - 1)(a + 1) \equiv 0 \pmod{p}$, and the latter is possible only when $a \equiv \pm 1 \pmod{p}$. Thus we find that

$$\prod_{a=2}^{p-2} a = \prod_{a} (a\overline{a}) \equiv 1 \pmod{p},$$

whence

$$\prod_{a=1}^{p-1} a \equiv (p-1) \equiv -1 \pmod{p}.$$

□

Notice that when m is a composite integer exceeding 4, then the product expansion of $(m-1)!$ necessarily contains two factors a and b with $m|ab$, whence one has $(m-1)! \equiv 0 \pmod{m}$. Wilson's Theorem therefore provides the world's worst primality test. On the other hand, the proof of Wilson's Theorem does motivate a proof of a criterion for the solubility of the congruence $x^2 \equiv -1 \pmod{p}$.

Theorem 4.15. *When $p = 2$, or when p is a prime number with $p \equiv 1 \pmod{4}$, the congruence $x^2 \equiv -1 \pmod{p}$ is soluble. When $p \equiv 3 \pmod{4}$, the latter congruence is not soluble.*

Proof. When $p = 2$, the conclusion is clear. Assume next that $p \equiv 1 \pmod{4}$, and write $r = (p-1)/2$ and $x = r!$. Then since r is even, one has

$$x^2 \equiv (1 \cdot 2 \cdot \dots \cdot r)((p-1) \cdot (p-2) \cdot \dots \cdot (p-r)) \equiv (p-1)! \equiv -1 \pmod{p}.$$

Thus, when $p \equiv 1 \pmod{4}$, the congruence $x^2 \equiv -1 \pmod{p}$ is indeed soluble. Suppose then that $p \equiv 3 \pmod{4}$. If it were possible that an integer x exists with $x^2 \equiv -1 \pmod{p}$, then one finds that $(x^2)^{(p-1)/2} \equiv (-1)^{(p-1)/2} \equiv -1 \pmod{p}$, yet by Fermat's Little Theorem, one has $(x^2)^{(p-1)/2} = x^{p-1} \equiv 1 \pmod{p}$ whenever $(x, p) = 1$. We therefore arrive at a contradiction, and this completes the proof of the theorem. □

The observation that -1 is not a square modulo p when $p \equiv 3 \pmod{4}$ can be exploited to provide simple irrationality proofs.

Theorem 4.16. *$\sqrt{2}$ is irrational.*

Proof. Suppose that $\sqrt{2}$ is rational, so there exist $x \in \mathbb{Z}$ and $y \in \mathbb{N}$ with $(x, y) = 1$ such that $(x/y)^2 = 2$. Then $x^2 = 2y^2$, so that in particular one has $x^2 \equiv 2y^2 \equiv -y^2 \pmod{3}$. But since $1^2 \equiv 2^2 \equiv 1 \pmod{3}$ and $0^2 \equiv 0 \pmod{3}$, it follows that the congruence $x^2 \equiv -y^2 \pmod{3}$ is soluble only when $3|x$ and $3|y$, and this contradicts the condition $(x, y) = 1$. We are therefore forced to conclude that $\sqrt{2}$ is irrational. □

Proof. (Novelty version) Suppose that $\sqrt{2}$ is rational, and let k be the smallest positive integer with $k\sqrt{2} \in \mathbb{Z}$. Then $k\sqrt{2} - k$ is a smaller such integer, contradicting the minimality of k and establishing the corollary. □

It may be worth expanding on the last line of this proof. Since $\sqrt{2}$ may be verified to lie between 1 and 2, and $k\sqrt{2}$ is supposed to be an integer, the number $k\sqrt{2} - k$ is a positive integer smaller than k. Moreover, since $(\sqrt{2})^2 = 2$ (it is here that the definition of $\sqrt{2}$ is used), one has $(k\sqrt{2} - k)\sqrt{2} = 2k - k\sqrt{2}$, and this is an integer because $k\sqrt{2}$ is again an integer. This is a proof that has been "rediscovered" many times (see, for example, T. Estermann, *The irrationality of* $\sqrt{2}$, Math. Gaz. (408) 59 (1975), 110).

5. The Chinese Remainder Theorem

We now seek to analyse the solubility of congruences by reinterpreting their solutions modulo a composite integer m in terms of related congruences modulo prime powers.

Theorem 5.1 (Chinese Remainder Theorem)**.** *Let $m_1, \dots, m_r$ denote positive integers with $(m_i, m_j) = 1$ for $i \neq j$. Also, let $a_1, \dots, a_r \in \mathbb{Z}$. Then the system of congruences*

$$x \equiv a_i \pmod{m_i} \quad (1 \leqslant i \leqslant r) \tag{5.1}$$

is soluble simultaneously for some integer x. If x_0 is any one such solution, then x is a solution of (5.1) *if and only if $x \equiv x_0 \pmod{m_1 m_2 \dots m_r}$.*

Proof. Let $m = m_1 m_2 \dots m_r$, and $n_j = m/m_j$ $(1 \leqslant j \leqslant r)$. Then for each j with $1 \leqslant j \leqslant r$ one has $(m_j, n_j) = 1$, whence by Theorem 4.13 there exists an integer b_j with $n_j b_j \equiv 1 \pmod{m_j}$. Moreover,

$$n_j b_j = \left(\frac{m_1 \dots m_r}{m_j m_i} b_j \right) m_i \equiv 0 \pmod{m_i}$$

whenever $i \neq j$. Then if we put $x_0 = n_1 b_1 a_1 + \dots + n_r b_r a_r$, we find that $x_0 \equiv n_i b_i a_i \equiv a_i \pmod{m_i}$ $(1 \leqslant i \leqslant r)$. Thus we may conclude that x_0 is a solution of (5.1).

In order to establish uniqueness, suppose that x and y are any two solutions of (5.1). Then one has $x \equiv y \pmod{m_i}$ $(1 \leqslant i \leqslant r)$ and $(m_i, m_j) = 1$ $(i \neq j)$. Then by Theorem 4.4(iii), it follows that $x \equiv y \pmod{[m_1, \dots, m_r]}$, and so $x \equiv y \pmod{m}$. □

Example 5.2. Find the set of solutions to the system of congruences

$$4x \equiv 1 \pmod 3, \quad x \equiv 2 \pmod 5, \quad 2x \equiv 5 \pmod 7.$$

We first convert this into a form where the leading coefficients are all 1. Thus, multiplying the final congruence through by 4 (the multiplicative inverse of 2 modulo 7), we obtain the equivalent system

$$x \equiv 1 \pmod 3, \quad x \equiv 2 \pmod 5, \quad x \equiv 6 \pmod 7.$$

We next put $m_1 = 3$, $m_2 = 5$, $m_3 = 7$, so that $(m_i, m_j) = 1$ for $i \neq j$. Define $m = 3 \cdot 5 \cdot 7 = 105$, and $n_1 = 105/3 = 35$, $n_2 = 105/5 = 21$, $n_3 = 105/7 = 15$. We compute integers b_j with $n_j b_j \equiv 1 \pmod{m_j}$ $(j = 1, 2, 3)$ by means of the Euclidean Algorithm (or directly, if the numbers are small enough). Thus we find that

$$\begin{aligned} &35 b_1 \equiv 1 \pmod 3 \Rightarrow 2 b_1 \equiv 1 \pmod 3 \Rightarrow b_1 \equiv 2 \pmod 3, \\ &21 b_2 \equiv 1 \pmod 5 \Rightarrow b_2 \equiv 1 \pmod 5, \\ &15 b_3 \equiv 1 \pmod 7 \Rightarrow b_3 \equiv 1 \pmod 7. \end{aligned}$$

So take

$$\begin{aligned} x_0 &= 35 \cdot 2 \cdot 1 + 21 \cdot 1 \cdot 2 + 15 \cdot 1 \cdot 6 \\ &= 70 + 42 + 90 = 202 \equiv 97 \pmod{105}. \end{aligned}$$

Then we find that $x_0 = 97$ satisfies the given congruences, and the complete set of solutions is given by $x = 97 + 105k$ $(k \in \mathbb{Z})$.

Example 5.3. Find the set of solutions, if any, to the system of congruences

$$x \equiv 1 \pmod{15}, \quad x \equiv 2 \pmod{35}.$$

In this example, the moduli of the two congruences are not coprime, since $(35,15)=5$. In order to determine whether or not the system is soluble, we therefore need to examine the underlying congruences, extracting as a modulus this greatest common divisor. Thus we find that any potential solution x of the system must satisfy

$$x\equiv 1 \pmod{15} \quad \Rightarrow \quad x\equiv 1 \pmod 3 \quad \text{and} \quad x\equiv 1 \pmod 5,$$

and at the same time

$$x\equiv 2 \pmod{35} \quad \Rightarrow \quad x\equiv 2 \pmod 5 \quad \text{and} \quad x\equiv 2 \pmod 7.$$

But then one has $x\equiv 1 \pmod 5$ and $x\equiv 2 \pmod 5$, two congruence conditions that are plainly incompatible. We may conclude then that there are no solutions of the simultaneous congruences $x\equiv 1 \pmod{15}$ and $x\equiv 2 \pmod{35}$.

We wish to investigate further the properties of the Euler totient, and so pause to introduce the concept of a multiplicative function.

Definition 5.4. (i) We say that a function $f:\mathbb{N}\rightarrow\mathbb{C}$ is an **arithmetical function**;
(ii) An arithmetical function f is said to be **multiplicative** if (a) f is not identically zero, and (b) whenever $(m,n)=1$, one has $f(mn)=f(m)f(n)$.

Note that if $f(n)$ is multiplicative, then necessarily one has $f(1)=1$ (Why?).

Theorem 5.5. *The function $\phi(n)$ is multiplicative. Thus, whenever $(m,n)=1$, one has $\phi(mn)=\phi(m)\phi(n)$. Moreover, if n has canonical prime factorisation $\prod_1^t p_i^{r_i}$, then*

$$\phi(n)=\prod_{i=1}^t p_i^{r_i-1}(p_i-1)=n\prod_{p|n}(1-1/p).$$

Proof. Let n and n' be natural numbers with $(n,n')=1$, and let a and a' run through the reduced residues modulo n, and modulo n' respectively. The total number of choices for a and a' is plainly $\phi(n)\phi(n')$. We examine the integer $an'+a'n$, and aim to show that this is a reduced residue modulo nn', and moreover that distinct choices for (a,a') yield distinct values of $an'+a'n \pmod{nn'}$. This shows that the number of reduced residues modulo nn' is at least as large as the number of pairs (a,a'), which is to say that one has $\phi(nn')\geqslant\phi(n)\phi(n')$.

Now, whenever $(a,n)=(a',n')=1$, one has

$$(an'+a'n,nn')|n'(an'+a'n,n)=n'(an',n)=n'(a,n)=n',$$

and likewise $(an'+a'n,nn')|n$, whence $(an'+a'n,nn')|(n,n')=1$. We therefore deduce that $(an'+a'n,nn')=1$, and so any integer of the shape $an'+a'n$, with $(a,n)=(a',n')=1$, is a reduced residue modulo nn'. But any two distinct numbers of the latter form are incongruent modulo nn', for if $(a_i,n)=(a_i',n')=1$ $(i=1,2)$, and

$$a_1n'+a_1'n\equiv a_2n'+a_2'n \pmod{nn'},$$

then

$$(a_1-a_2)n'\equiv 0 \pmod n \quad \Rightarrow \quad a_1\equiv a_2 \pmod n,$$

and similarly $a_1'\equiv a_2' \pmod{n'}$. Thus we obtain $a_1=a_2$ and $a_1'=a_2'$. Distinct choices for (a,a') do indeed lead to distinct values of $an'+a'n \pmod{nn'}$, therefore, and we have achieved the objectives described in the first paragraph of our proof.

We next seek to establish that whenever $(b, nn') = 1$, then there exist reduced residues a modulo n and a' modulo n' with $b \equiv an' + a'n \pmod{nn'}$. But $(n, n') = 1$, so by the Euclidean Algorithm, there exist integers m and m' with $mn' + m'n = 1$. Now $(m, n) = (m', n') = 1$, and so $(bm, n) = (bm', n') = 1$, and thus there exist integers a and a' with $(a, n) = (a', n') = 1$ satisfying $an' + a'n = b$, namely $a = bm$ and $a' = bm'$. Distinct choices for b generate distinct choices for a and a' modulo n and modulo n', respectively. For if b_i generates (a_i, a_i') for $i = 1, 2$, and $a_1 \equiv a_2 \pmod{n}$ and $a_1' \equiv a_2' \pmod{n'}$, then $b_1 m = a_1 \equiv a_2 = b_2 m \pmod{n}$, whence $b_1 \equiv b_2 \pmod{n}$, since $(m, n) = 1$. Similarly, one has $b_1 \equiv b_2 \pmod{n'}$, and thus $b_1 \equiv b_2 \pmod{nn'}$, which shows that $b_1 = b_2$. It therefore follows that the number of pairs (a, a') with a a reduced residue modulo n, and a' a reduced residue modulo n', cannot be smaller than the number of reduced residues modulo nn'. This establishes that $\phi(nn') \leqslant \phi(n)\phi(n')$. Together with the inequality $\phi(nn') \geqslant \phi(n)\phi(n')$ that we established earlier, this yields the relation $\phi(nn') = \phi(n)\phi(n')$ whenever $(n, n') = 1$. Then the Euler totient is indeed a multiplicative function.

In order to complete the proof of the theorem, we observe next that when p is a prime number, one has $\phi(p^r) = p^r - p^{r-1}$, since the total number of residues modulo p^r is p^r, of which precisely the p^{r-1} divisible by p are not reduced. In this way, the final assertions of the theorem follow by making use of the multiplicative property of $\phi(\cdot)$. □

Useful properties of $\phi(n)$ that will be employed later stem easily from its multiplicative property. Before establishing one such property, we establish a general result for multiplicative functions.

Lemma 5.6. *Suppose that $f(n)$ is multiplicative, and define $g(n) = \sum_{d|n} f(d)$. Then $g(n)$ is a multiplicative function.*

Proof. Suppose that n and m are natural numbers with $(n, m) = 1$, and suppose that $d|mn$. Write $d_1 = (d, m)$ and $d_2 = (d, n)$. Then $d = d_1 d_2$ and $(d_1, d_2) = 1$. Thus we obtain

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

whence $g(mn) = g(m)g(n)$. This completes the proof that g is multiplicative. □

Corollary 5.7. *One has $\sum_{d|n} \phi(d) = n$.*

Proof. Observe that for each prime number p, and every natural number r, one has

$$\sum_{d|p^r} \phi(d) = \sum_{h=0}^{r} \phi(p^h) = 1 + \sum_{h=1}^{r} (p^h - p^{h-1}) = p^r.$$

Thus, owing to the multiplicative property of ϕ established in Theorem 5.5, it follows from Lemma 5.6 that $\sum_{d|n} \phi(d)$ is a multiplicative function of n, whence

$$\sum_{d|n} \phi(d) = \prod_{p^r \| n} \left(\sum_{d|p^r} \phi(d) \right) = \prod_{p^r \| n} p^r = n.$$

□

To conclude this section, we examine the set of solutions of a polynomial congruence.

Definition 5.8. Let $f(x) \in \mathbb{Z}[x]$, and suppose that $r_1, \ldots, r_m$ is a complete residue system modulo m. Then we say that the **number of solutions** of the congruence $f(x) \equiv 0 \pmod{m}$ is the number of residues r_i with $f(r_i) \equiv 0 \pmod{m}$.

Definition 5.9. Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a polynomial with integral coefficients. Let j be the largest integer with $m \nmid a_j$. Then we say that the **degree** of f modulo m is j. If $m | a_j$ for every j, then the degree of f is undefined.

Theorem 5.10. *Suppose that $f(x) \in \mathbb{Z}[x]$, and denote by $N_f(m)$ the number of solutions of the congruence $f(x) \equiv 0 \pmod{m}$. Then $N_f(m)$ is a multiplicative function of m, and*

$$N_f(m) = \prod_{p^r \| m} N_f(p^r).$$

Proof. Suppose that m_1 and m_2 are natural numbers with $m = m_1 m_2$ and $(m_1, m_2) = 1$. Whenever $f(a) \equiv 0 \pmod{m}$, one has also $f(a) \equiv 0 \pmod{m_1}$ and $f(a) \equiv 0 \pmod{m_2}$. Then if $\{r_1, \ldots, r_{m_1}\}$ and $\{s_1, \ldots, s_{m_2}\}$ are complete residue systems modulo m_1 and m_2, respectively, one finds that for each integer a with $f(a) \equiv 0 \pmod{m}$ belonging to a complete residue system modulo m, there exist unique r_i and s_j with $f(r_i) \equiv 0 \pmod{m_1}$ and $f(s_j) \equiv 0 \pmod{m_2}$. Moreover, the residue a modulo $m_1 m_2$ satisfying $a \equiv r_i \pmod{m_1}$ and $a \equiv s_j \pmod{m_2}$ is uniquely defined, as a consequence of the Chinese Remainder Theorem. Thus there is an injective map from the set of solutions modulo m to the set of pairs of solutions modulo m_1 and m_2.

In the other direction, whenever there exist residues r_i and s_j with $f(r_i) \equiv 0 \pmod{m_1}$ and $f(s_j) \equiv 0 \pmod{m_2}$, then by the Chinese Remainder Theorem there exists an integer a with $a \equiv r_i \pmod{m_1}$ and $a \equiv s_j \pmod{m_2}$ such that $f(a) \equiv 0 \pmod{m_i}$ $(i = 1, 2)$, and moreover the integer a uniquely defines r_i modulo m_1 and s_j modulo m_2. But since $(m_1, m_2) = 1$, it follows that $f(a) \equiv 0 \pmod{m_1 m_2}$, whence $f(a) \equiv 0 \pmod{m}$. There is therefore an injective map from pairs of solutions (r_i, s_j) modulo m_1 and m_2 respectively, to solutions modulo m.

Collecting together the above conclusions, we find that the solutions modulo m, and pairs of solutions modulo m_1 and m_2, are in bijective correspondence, whence $N_f(m) = N_f(m_1)N_f(m_2)$ whenever $(m_1, m_2) = 1$. The desired conclusion now follows on considering the prime factorisation of m. □

6. Public-Key Cryptography: the RSA cryptosystem [Non-examinable]

Suppose that Alice wishes to securely send a message to Bob, avoiding Eve malevolently deciphering this message. Say the message is:

"Do not spill the beans"

How do we achieve secure communication? We will provide a sketch of the RSA cryptosystem, described by Rivest, Shamir and Adleman in 1977, and patented in the USA in 1983[1].

[1] See R. L. Rivest, A. Shamir and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems.* Comm. ACM 21 (1978), 120–126. There is an interesting history to the RSA cryptosystem: Cliff Cocks at GCHQ devised such a method in 1973, though owing to the secrecy of GCHQ operations, this information became publicly available only in 1997.

Step 1: Bob publishes a pair of integers N, r (*the Public Key*).
As the latter name suggests, these integers are in the public domain and can be used by anyone (including Alice) to communicate securely with Bob. Bob obtains these two integers as follows. He picks two large primes p and q in an essentially random manner, with $p \neq q$. In practice, one should choose these primes to have 150 - 200 digits, but in order to illustrate ideas, we'll take $p = 257$ and $q = 8191$. The number N is then taken to be $pq = 2\,105\,087$. Bob keeps the identity of these two primes secret. It is only the product N which is put into the public domain. The second integer r is chosen by Bob to be a natural number coprime to $\phi(N)$ that is not too small. Notice that since Bob knows the prime factorisation of N, he is able to compute $\phi(N) = (p-1)(q-1)$ quickly, and hence obtain a suitable integer r by trial and error using the Euclidean Algorithm. In this discussion we take $r = 139$. Thus the Public Key is $(2\,105\,087, 139)$.
Step 2: Alice now needs to code her message into a numerical expression in a standard manner. Obvious choices for a suitable scheme include the ASCII scheme (which also offers the possibility of encoding punctuation symbols and so on). For simplicity, we'll encode "A" as "01", "B" as "02", ..., "Z" as "26", and "space" as "27". Thus Alice's message is encoded as

$$04|15|27|14|15|20|27|19|16|09|12|12|27|20|08|05|27|02|05|01|14|19$$

Alice now needs to break this string of numbers up into smaller substrings that can be encrypted using Bob's public key. Since N has seven digits, this entails breaking the message into substrings having six digits apiece. The message becomes $a_1a_2\dots a_8$, where

$$\begin{aligned} a_1 &= 041527, & a_2 &= 141520, & a_3 &= 271916, & a_4 &= 091212,\\ a_5 &= 272008, & a_6 &= 052702, & a_7 &= 050114, & a_8 &= 198888. \end{aligned}$$

Notice here that the last substring a_8 has been padded with the digit 8 to boost it to the correct length. The question of appropriate padding schemes is one of some subtlety if security is to be preserved. Alice now computes the residues $b_i \equiv a_i^r \pmod{N}$ efficiently by using the "divide-and-conquer" algorithm for $1 \leqslant i \leqslant 8$.
Step 3: Alice may now send Bob the message $b_1b_2\dots b_8$, where

$$\begin{aligned} b_1 &= 0994340, & b_2 &= 0128098, & b_3 &= 1608212, & b_4 &= 0600447,\\ b_5 &= 1096537, & b_6 &= 0305539, & b_7 &= 0137494, & b_8 &= 1528105. \end{aligned}$$

Step 4: Bob now needs to decode the message, but because he knows the two primes p and q for which $N = pq$, he can compute

$$\phi(N) = \phi(pq) = (p-1)(q-1) = 2\,096\,640.$$

Eve cannot compute $\phi(N)$ easily without knowing p and q. Thus Bob can find an integer s such that $sr \equiv 1 \pmod{\phi(N)}$, say $sr - 1 = -k\phi(N)$, for a suitable integer k. One can compute this number s by using the Euclidean Algorithm to solve the linear equation $xr + \phi(N)y = 1$. Thus Bob solves the equation

$$139s + 2\,096\,640t = 1.$$

One may verify that $(s,t) = (1\,689\,379, -112)$ solves this equation. Of course, Bob only needs to solve this equation once so long as he stores the Private Key (N, s) in a secure place. Now Bob computes the residues $b_i^s \pmod{N}$ for $1 \leqslant i \leqslant 8$ in order to recover the original message $a_1 a_2 \dots a_8$, that is

$$04|15|27|14|15|20|27|19|16|09|12|12|27|20|08|05|27|02|05|01|14|19|88|88$$

complete with padding at the end. Bob may then of course decode the message sing the transparent coding scheme to obtain "DO NOT SPILL THE BEANS".

Observation 6.1. *One has* $b_i^s \equiv a_i \pmod{N}$ *for each* i.

Proof. Suppose first that $(a_i, N) = 1$. Then it follows from Euler's Theorem that

$$b_i^s \equiv a_i^{rs} \equiv a_i^{rs}(a_i^{\phi(N)})^k = a_i^{rs+k\phi(N)} \equiv a_i^1 \pmod{N}.$$

Moreover, since $N = pq$, it follows that when $(a_i, N) \neq 1$, then one has $(a_i, N) = p$, q or pq. In the latter case, we have $a_i = pq = N$, and then the conclusion is trivial. Suppose then that $(a_i, N) = p$, so that $p|a_i$ and $(a_i, q) = 1$. In this situation the former condition yields

$$b_i^s \equiv a_i^{rs} \equiv 0 \equiv a_i \pmod{p},$$

and in view of Fermat's Little theorem, the latter yields

$$b_i^s = a_i^{rs}(a_i^{q-1})^{k(p-1)} = a_i^{rs+k\phi(N)} \equiv a_i^1 \pmod{q}.$$

Thus $b_i^s \equiv a_i \pmod{p}$ and $b_i^s \equiv a_i \pmod{q}$, whence $b_i^s \equiv a \pmod{pq}$. The situation in which $(a_i, N) = q$ may be analysed in like manner, and so this completes the proof. □

One could argue, of course, that to send a message that contains a common factor with N that yields the prime factorisation of N would be foolish, and something to be avoided by using a suitable padding scheme.

It remains to discuss the feasibility and security of this cryptosystem. The first observation to make is that all of the operations required to make use of the RSA cryptosystem are fast. The application of the Euclidean Algorithm, and the operation of taking powers modulo N, have running time $O(\log N)$ arithmetic operations. This is proportional to the number of digits in N. Second, we need to have available plenty of large prime numbers (p and q) in order to derive good public keys. Fortunately, there are relatively fast primality tests available. A probabilistic test is available with running time polynomial in $\log n$ that can discern, provably, that a number n is composite. For the numbers that survive this test, the Adleman-Pomerance-Rumely test can establish primality, or compositeness, provably in deterministic time $O((\log n)^{c \log\log\log n})$, which is close to polynomial in $\log n$. More recently, Agrawal, Kayal and Saxena have devised an algorithm that has running time polynomial in $\log n$. Finally, the security of the RSA cryptosystem depends on the difficulty of factoring large integers. The naive factorisation algorithm supplies a factorisation of a composite integer in running time $O(\sqrt{n})$ arithmetic operations. The fastest available factorisation algorithm for very large integers is the Number Field Sieve, with running time $\exp(c(\log n)^{1/3}(\log\log n)^{2/3})$ arithmetic operations to factor a large integer n, wherein c is a suitably large positive constant. This is much larger than polynomial in $\log n$. If a quantum computer can be built, then Shor's Quantum Algorithm would factor integers n in a time polynomial in $\log n$, and would constitute a threat to the RSA cryptosystem.

Pollard's rho-method [Non-examinable].
We briefly explore a factorisation algorithm that has running time significantly faster than the naive one. Note first that on observing that a composite number $n = n_1 n_2$ has one factor at least smaller than $\sqrt{n}$, it is apparent that simply by testing each possible factor smaller than $\sqrt{n}$, one obtains a factorisation algorithm with running time $O(\sqrt{n})$. Pollard's rho-method, which we now describe, has expected running time about $O(n^{1/4})$.

Suppose that n is a large composite number with smallest prime divisor p. Choose k to be large compared to $\sqrt{p}$, say $k = 10n^{1/4}$, and choose k integers $u_1, \dots, u_k$ by some "random" (or rather, quasi-random) process. Then with high probability, the u_i are distinct modulo n. The probability that two u_i are mutually congruent modulo p is $1 - \pi$, where π is the probability that they are all distinct. But

$$\begin{aligned}\pi &\approx \left(1 - \frac{1}{p}\right)\left(1 - \frac{2}{p}\right)\dots\left(1 - \frac{k-1}{p}\right)\\ &\approx \exp\left(-\frac{1}{p} - \frac{2}{p} - \dots - \frac{k-1}{p}\right)\\ &\approx \exp\left(-\frac{k(k-1)}{2p}\right).\end{aligned}$$

But $k = 10n^{1/4} \geqslant 10p^{1/2}$, so π is no larger than about e^{-50}, which is microscopic. Thus, almost certainly, one finds that there are two numbers u_i and u_j with $1 < (u_i - u_j, n) < n$, and hence we obtain a non-trivial factor of n.

We must now obtain a suitable pseudo-random sequence (u_i) with which to put this idea into effect. It transpires that when $c \neq 0, -2$, the sequence generated with some initial good seed u_0, and defined for $i \geqslant 1$ via the relation $u_{i+1} \equiv u_i^2 + c \pmod{n}$, is pseudo-random. Notice here that we could omit the reduction modulo n in the definition, but that taking the numerically least residue offers computational advantages.

Example 6.2. Consider the integer $n = 78\,667$. Make use of the pseudo-random sequence defined by $u_0 = 3$, $u_{i+1} = u_i^2 - 1 \pmod{n}$ to obtain a factorisation of n.

One may compute that the sequence $\{u_i \pmod{n}\}$ is

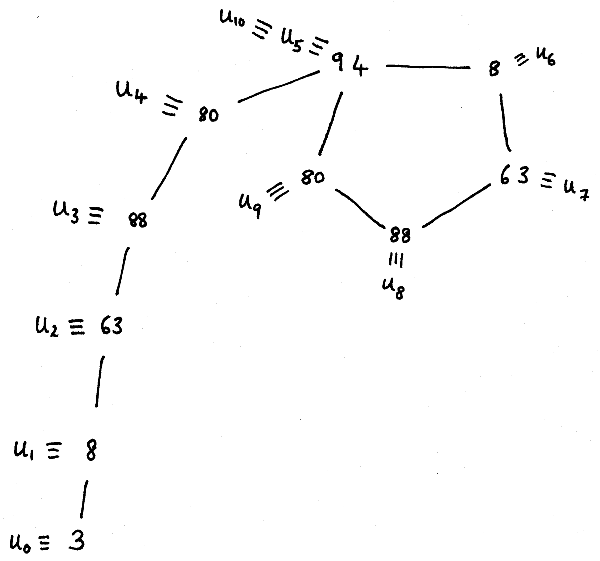
$$\{3,\ 8,\ 63,\ 3\,968,\ 11\,623,\ 22\,889,\ 62\,767,\ 52\,928,\ 41\,313,\ 4\,736,\ 9\,600, \dots\},$$

and hence $(u_{10} - u_5, n) = 97$, giving $78\,667 = 97 \cdot 811$.

This algorithm is only fast provided that we can detect the mutual congruences efficiently. But using the polynomial pseudo-random generator, one can proceed as follows. If $u_i \equiv u_j \pmod{d}$ for some integer d with $d|n$, then $u_{i+1} \equiv u_i^2 + c \equiv u_j^2 + c \equiv u_{j+1} \pmod{d}$, and hence the sequence (u_i) is ultimately periodic modulo d, with period dividing $j - i$. Put $r = j - i$. Then $u_s \equiv u_t \pmod{d}$ whenever $s \equiv t \pmod{r}$ and $s \geqslant i$, $t \geqslant i$. Let s be the least multiple of r exceeding $i - 1$, and take $t = 2s$. Then $u_s \equiv u_{2s} \pmod{d}$. Consequently, amongst the numbers $u_{2s} - u_s$, we expect to find one with $1 < (u_{2s} - u_s, n) < n$, with $s \leqslant 10n^{1/4}$. One can of course compute the pair (u_s, u_{2s}) for successive values of s relatively efficiently. One has $(u_{s+1}, u_{2s+2}) \equiv (u_s^2 + c, (u_{2s}^2 + c)^2 + c) \pmod{n}$, and so the expected running time required to find a factorisation is $O(n^{1/4})$. The name of the algorithm then derives from the shape of a tree of the iterates, ultimately periodic modulo n (see Figure 1).

Figure 1.

u_i (mod 97)



7. Polynomial congruences to prime modulus

At this point we know that the number of solutions of a polynomial congruence modulo m is a multiplicative function of m, and thus it suffices to consider congruences modulo prime powers. We begin by investigating congruences modulo p, for prime numbers p.

Theorem 7.1 (Lagrange). *Let $f(x) \in \mathbb{Z}[x]$ have degree n (modulo p), with $n \geqslant 1$. Then the congruence $f(x) \equiv 0 \pmod{p}$ has at most n solutions.*

Proof. The situation when $n = 1$ is clear, since we then have a linear equation to solve. Suppose then that $n \geqslant 2$, and that the conclusion of the theorem holds for all degrees smaller than n. Let $f(x) \in \mathbb{Z}[x]$ have degree n modulo p. Either $f(x)$ has no zeros modulo p, or else there exists at least one zero, say $x = a$. Let $g_a(x)$ be defined by means of the relation $f(x) - f(a) = (x - a)g_a(x)$. By considering the polynomials $(x^m - a^m)/(x - a)$, it is apparent that $g_a(x) \in \mathbb{Z}[x]$ and that $g_a(x)$ has degree $n - 1$ modulo p. It follows that whenever $f(x) \equiv 0 \pmod{p}$, one has either $x \equiv a \pmod{p}$, or $g_a(x) \equiv 0 \pmod{p}$. But by our inductive hypothesis, the number of zeros of $g_a(x) \pmod{p}$ is at most $\deg g_a = n - 1$, and hence the number of zeros of $f(x) \pmod{p}$ is at most $1 + (n - 1) = n$. The desired conclusion therefore follows by induction. □

Note that from what we have already discussed, it follows that the set of residues modulo p, namely $\mathbb{Z}/p\mathbb{Z}$, forms a field under addition and multiplication modulo p. Then the above theorem is immediate from the standard properties of fields.

Example 7.2. (i) It follows from Lagrange's Theorem that the congruence $x^2 + 1 \equiv 0 \pmod{p}$ has at most 2 solutions for any prime p. From Theorem 4.15, meanwhile, we know that this congruence has precisely 2 solutions when $p \equiv 1 \pmod{4}$, and 0 solutions when $p \equiv 3 \pmod{4}$.

(ii) It follows from Lagrange's Theorem that the congruence $x^p - x + 1 \equiv 0 \pmod{p}$ has at most p solutions modulo p. In fact this congruence has no solutions for any prime p, as a consequence of Fermat's Little Theorem (see homework sheet 4).
(iii) There is no analogue of Lagrange's Theorem for composite moduli. Consider for example the congruence $x^2 \equiv 1 \pmod{8}$. This is a congruence of degree 2, yet has 4 distinct solutions 1, 3, 5 and 7 modulo 8.

Continuing the inductive argument of the proof of the last theorem, we find that whenever $a_1, \dots, a_n$ are zeros of a polynomial $f(x) \pmod{p}$, counted with multiplicity, and $f(x)$ has degree n modulo p, then there exists a non-zero residue $\alpha \pmod{p}$ with the property that

$$f(x) \equiv \alpha(x - a_1)(x - a_2)\dots(x - a_n) \pmod{p}.$$

In particular, by Lagrange's Theorem, the congruence $x^{p-1} \equiv 1 \pmod{p}$ has at most $p - 1$ solutions modulo p, and it follows from Fermat's Little Theorem that these are $x = 1, 2, \dots, p - 1$. Thus one obtains the relation

$$x^{p-1} - 1 \equiv (x - 1)(x - 2)\dots(x - p + 1) \pmod{p}.$$

Comparing coefficients of powers of x, we find from the constant coefficient in this relation that $(p - 1)! \equiv -1 \pmod{p}$. Moreover, on writing $1/n$ for the multiplicative inverse of n modulo p, it follows by comparing the coefficients of x that

$$(p-1)!\left(\frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{p-1}\right) \equiv 0 \pmod{p},$$

whence

$$1 + \frac{1}{2} + \dots + \frac{1}{p-1} \equiv 0 \pmod{p}.$$

More is true. One can prove (Wolstenholme's Theorem) that

$$1 + \frac{1}{2} + \dots + \frac{1}{p-1} \equiv 0 \pmod{p^2}.$$

Corollary 7.3. *Whenever $d|(p - 1)$, the congruence $x^d \equiv 1 \pmod{p}$ has precisely d solutions modulo p.*

Proof. Suppose that $d|(p-1)$. Then there exists a polynomial $g(x) \in \mathbb{Z}[x]$ with $x^{p-1} - 1 = (x^d)^{(p-1)/d} - 1 = (x^d - 1)g(x)$. But the degree of g is $p - 1 - d$, and so by Lagrange's Theorem the congruence $g(x) \equiv 0 \pmod{p}$ has at most $p - 1 - d$ solutions modulo p. Then since $x^{p-1} - 1$ has precisely $p - 1$ zeros modulo p, we see from the above relation that $x^d - 1$ has at least d zeros modulo p. But Lagrange's Theorem shows that the latter polynomial has at most d zeros modulo p, and thus we see that it has precisely d zeros modulo p. This completes the proof of the theorem. □

8. Congruences to prime power moduli

Although there is no analogue of Lagrange's Theorem for prime power moduli, there is an algorithm for determining when a solution modulo p generates solutions to higher power moduli. The motivation comes from Newton's method for approximating roots over the real numbers. We first present a motivating example.

Example 8.1. Solve the congruence $x^3 + x + 4 \equiv 0 \pmod{7^3}$.

(I) We first solve the corresponding congruence modulo 7, since any solution x modulo 7^3 must also satisfy $x^3+x+4 \equiv 0 \pmod{7}$. By an exhaustive search (try $x=0,1,2,...,6$), we find that the only solution is $x \equiv 2 \pmod{7}$.

(II) Next, we try to solve the corresponding congruence modulo 7^2, since any solution x modulo 7^3 must also satisfy $x^3+x+4 \equiv 0 \pmod{7^2}$. But such solutions must also satisfy the corresponding solution modulo 7, so $x \equiv 2 \pmod{7}$. Then we put $x = 2+7y$ and substitute. We need to solve

$$(2+7y)^3+(2+7y)+4 \equiv 0 \pmod{7^2}.$$

Notice that when we use the Binomial Theorem to expand the cube, any terms involving 7^2 or 7^3 can be ignored. Thus we need to solve

$$(2^3+3\cdot 2^2\cdot 7y)+(2+7y)+4 = 14+13\cdot 7y \equiv 0 \pmod{7^2},$$

or equivalently,

$$13y+2 \equiv -y+2 \equiv 0 \pmod{7}.$$

Then we put $y=2$ and find that $x=2+7y=16$ satisfies the congruence $x^3+x+4 \equiv 0 \pmod{7^2}$.

(III) We can now repeat the previous strategy (and in fact, we can repeat this as many times as necessary). So we substitute $x=16+7^2z$ and solve for z to obtain a solution modulo 7^3. Thus we need to solve

$$(16+7^2z)^3+(16+7^2z)+4 \equiv (16^3+3\cdot 16^2\cdot 7^2z)+(16+7^2z)+4 \equiv 0 \pmod{7^3}.$$

But 16^3+16+4 is divisible by 7^2 (why do we know this?), and in fact is equal to $84\cdot 7^2$. Then we need to solve

$$84\cdot 7^2+(3\cdot 16^2+1)\cdot 7^2z \equiv 0 \pmod{7^3},$$

which is equivalent to

$$(3\cdot 16^2+1)z+84 \equiv 0 \pmod{7},$$

or $13z \equiv 0 \pmod{7}$. So we put $z=0$, and find that $x \equiv 16 \pmod{7^3}$ solves $x^3+x+4 \equiv 0 \pmod{7^3}$.

Theorem 8.2 (Hensel's Lemma). *Let $f(x)\in\mathbb{Z}[x]$. Suppose that $f(a)\equiv 0 \pmod{p^j}$, and that $p^\tau \| f'(a)$. Then if $j \geqslant 2\tau+1$, it follows that:*
(i) whenever $b \equiv a \pmod{p^{j-\tau}}$, one has $f(b)\equiv f(a) \pmod{p^j}$ and $p^\tau\|f'(b)$;
(ii) there is a unique residue $t \pmod{p}$ such that $f(a+tp^{j-\tau})\equiv 0 \pmod{p^{j+1}}$.

Proof. First consider part (i) of the theorem. Let the integers a and b satisfy the hypotheses of the statement of the theorem, and define the integer h by means of the relation $b-a=hp^{j-\tau}$. Then by the binomial theorem, which for polynomials we can interpret as a version of Taylor's theorem, it follows that

$$f(b)=f(a+hp^{j-\tau})=f(a)+hp^{j-\tau}f'(a)+\frac{1}{2!}f''(a)(hp^{j-\tau})^2+\dots.$$

Despite the presence of reciprocals of factorials, the coefficients in the above Taylor expansion are necessarily integral. It is for this purpose that we regard the Taylor expansion as an application of the binomial theorem, in which each monomial x^m occuring in $f(x)$

is expanded individually. Thus the third and higher terms in the above expansion are all divisible by $p^{2(j-\tau)}$. But $j \geqslant 2\tau+1$, whence $2(j-\tau) \geqslant j+(j-2\tau) > j$, and so

$$f(b) \equiv f(a) + hp^{j-\tau} f'(a) \pmod{p^j}.$$

Since $p^\tau | f'(a)$, the latter shows that $f(b) \equiv f(a) \pmod{p^j}$. Moreover, applying the binomial theorem in like manner, one finds that

$$\begin{aligned} f'(b) = f'(a+hp^{j-\tau}) &\equiv f'(a) \pmod{p^{j-\tau}} \\ &\equiv f'(a) \pmod{p^{\tau+1}}, \end{aligned}$$

since $j-\tau \geqslant \tau+1$. Then since $p^\tau \| f'(a)$, one obtains $p^\tau \| f'(b)$, and this completes the proof of the first part of the theorem.

Now we turn to the second part of the theorem. Since $p^\tau \| f'(a)$, we may write $f'(a) = gp^\tau$ for a suitable integer g with $(g,p)=1$. Let $\overline{g}$ be any integer with $g\overline{g} \equiv 1 \pmod{p}$, and write $a' = a - \overline{g} f(a) p^{-\tau}$. Then an application of the binomial theorem on this occasion supplies the congruence

$$f(a') = f(a - \overline{g} f(a) p^{-\tau}) \equiv f(a) - p^{-\tau} f(a) \overline{g} f'(a) \pmod{p^{2(j-\tau)}},$$

since $p^{-\tau}\overline{g} f(a) \equiv 0 \pmod{p^{j-\tau}}$ and $j \geqslant \tau+1$. But $2(j-\tau) \geqslant j+1$, and thus

$$f(a') \equiv f(a) - (p^{-\tau} f(a) \overline{g})(gp^\tau) \equiv f(a) - f(a) g\overline{g} \equiv 0 \pmod{p^{j+1}}.$$

So there exists an integer t with $f(a+tp^{j-\tau}) \equiv 0 \pmod{p^{j+1}}$, and indeed one may take $t \equiv -p^{-j} f(a)(p^{-\tau} f'(a))^{-1} \pmod{p}$.

In order to establish the uniqueness of the integer t, suppose, if possible, that two such integers t_1 and t_2 exist. Then one has

$$f(a+t_1 p^{j-\tau}) \equiv 0 \equiv f(a+t_2 p^{j-\tau}) \pmod{p^{j+1}},$$

whence by the binomial theorem, as above, one obtains

$$f(a) + t_1 p^{j-\tau} f'(a) \equiv f(a) + t_2 p^{j-\tau} f'(a) \pmod{p^{j+1}}.$$

Thus $t_1 f'(a) \equiv t_2 f'(a) \pmod{p^{\tau+1}}$. Since $p^\tau \| f'(a)$, we obtain $t_1 \equiv t_2 \pmod{p}$. This establishes the uniqueness of t modulo p, completing our proof. □

Example 8.3. Let $f(x) = x^2+1$. Find the solutions of the congruence $f(x) \equiv 0 \pmod{5^4}$.

Observe that the congruence $x^2+1 \equiv 0 \pmod{5}$ has the solutions $x \equiv \pm 2 \pmod{5}$ (note that there are at most 2 solutions modulo 5, by Lagrange's theorem). Consider first the solution $x_0 = 2$ of the latter congruence. One finds that $f'(x_0) = 2x_0 \equiv -1 \pmod{5}$. It follows that $5^0 \| f'(x_0)$, and since $f(x_0) = 5 \equiv 0 \pmod{5}$, we may apply Hensel's iteration to find integers x_n $(n \geqslant 1)$ with $f(x_n) \equiv 0 \pmod{5^n}$. We obtain

$$\begin{aligned} x_1 &\equiv x_0 - \frac{f(x_0)}{f'(x_0)} \equiv 2 - \frac{5}{-1} \equiv 7 \pmod{5^2}, \\ x_2 &\equiv 7 - \frac{50}{14} \equiv 7 - \frac{50}{-1} \equiv 57 \pmod{5^3} \\ x_3 &\equiv 57 - \frac{3\,250}{114} \equiv 57 - \frac{3\,250}{-1} \equiv 3307 \equiv 182 \pmod{5^4}. \end{aligned}$$

Thus $x = 182$ provides a solution of the congruence $x^2 + 1 \equiv 0 \pmod{5^4}$. Proceeding similarly, one may lift the alternate solution $x = -2$ to the congruence $x^2+1 \equiv 0 \pmod 5$ to obtain the solution $x = -182 \pmod{5^4}$. Note that in each instance, the lifting process provided by Hensel's lemma led to a unique residue modulo 5^4 corresponding to each starting solution modulo 5.

Example 8.4. Let $f(x) = x^2 - 4x + 13$. Find all of the solutions of the congruence $f(x) \equiv 0 \pmod{3^4}$.

Notice that

$$x^2 - 4x + 13 \equiv x^2 + 2x + 1 \equiv (x + 1)^2 \pmod 3,$$

and hence $x \equiv -1 \pmod 3$ is the only solution of the congruence $f(x) \equiv 0 \pmod 3$. Next, since $f'(x) = 2x - 4$, we find that $3\|f'(-1)$, and so in order to apply Hensel's lemma, we must determine all of the solutions of the congruence $f(x) \equiv 0 \pmod{3^3}$. We proceed systematically.

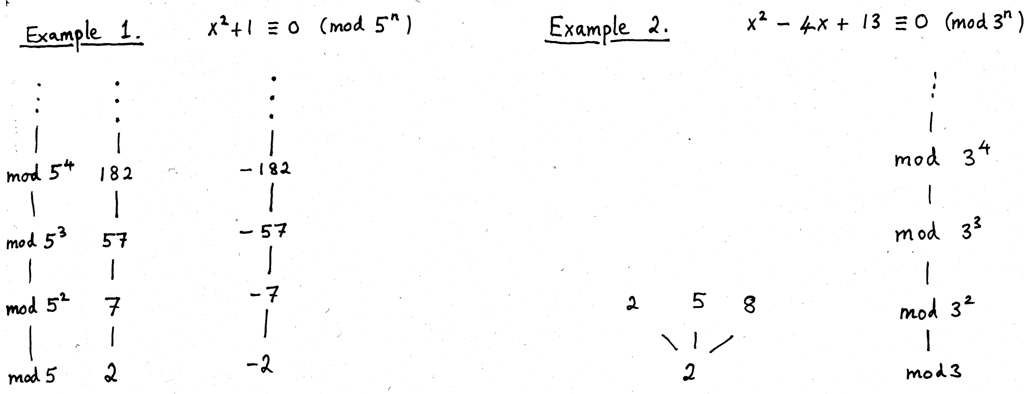
(i) Observe first that any solutions satisfy $x \equiv 2 \pmod 3$, and so any solution x must satisfy $x \equiv 2$, 5 or 8 modulo 9. One may verify that all three residue classes satisfy $f(x) \equiv 0 \pmod 9$.

(ii) Next we consider all residues modulo 27 satisfying $x \equiv 2$, 5 or 8 modulo 9, and find that none of these (there are 9 such residues) provide solutions of $f(x) \equiv 0 \pmod{27}$.

So there are no solutions to the congruence $x^2 - 4x + 13 \equiv 0 \pmod{3^4}$.

See Figure 2 for a pictorial representation of the lifting process in these two examples.

Figure 2.



Some concluding observations may be of assistance:

(i) Hensel's lemma allows one to lift repeatedly. Thus, whenever

$$f(a) \equiv 0 \pmod{p^j} \text{ and } p^\tau\|f'(a) \text{ with } j \geqslant 2\tau + 1$$

then there exists a unique residue t modulo p such that, with $a' = a + tp^{j-\tau}$,

$$f(a') \equiv 0 \pmod{p^j} \text{ and } p^\tau\|f'(a') \text{ with } j + 1 \geqslant 2\tau + 1,$$

and then we are set up to repeat this process.

(ii) Notice that in Hensel's lemma, the residue t modulo p is unique, and given by

$$t \equiv -(p^{-j}f(a))(p^{-\tau}f'(a))^{-1} \pmod p,$$

so one only needs to compute $(p^{-\tau}f'(a))^{-1}$ modulo p. Moreover, $p^{-\tau}f'(a') \equiv p^{-\tau}f'(a) \pmod{p}$, so our initial inverse computation remains valid for subsequent lifting processes.
(iii) If $f(a) \equiv 0 \pmod{p^j}$ and $p^\tau \| f'(a)$ and $j \geqslant 2\tau + 1$, then

$$f(a + hp^{j-\tau}) \equiv f(a) \equiv 0 \pmod{p^j}.$$

So there are p^τ solutions of $f(x) \equiv 0 \pmod{p^j}$ corresponding to the single solution $x \equiv a \pmod{p^j}$, namely $a + hp^{j-\tau}$ with $0 \leqslant h \leqslant p^\tau$.

A sketch of the p-adic numbers (non-examinable). Let us begin by recalling how the real numbers $\mathbb{R}$ are defined starting from $\mathbb{Q}$. One begins with two ingredients: (i) the set of rational numbers $\mathbb{Q}$, and (ii) the ordinary absolute value $|\cdot|$. Now consider the set of Cauchy sequences in $\mathbb{Q}$, that is, the set of sequences $(a_n)_{n=1}^\infty$ satisfying the property that whenever $\varepsilon > 0$, there exists $N = N(\varepsilon)$ such that whenever $n > m > N(\varepsilon)$, one has $|a_n - a_m| < \varepsilon$. Define

$$\mathcal{R} = \{(a_n)_{n=1}^\infty : a_n \in \mathbb{Q} \text{ for each } n, \text{ and } (a_n) \text{ is a Cauchy sequence}\}.$$

One can show that $\mathcal{R}$ forms a ring under addition and multiplication defined coordinatewise in the obvious fashion. Now identify two Cauchy sequences (a_n) and (b_n) when $\lim_{n\to\infty} |a_n - b_n| = 0$. Modulo this equivalence, we may label Cauchy sequences, say $\alpha = (a_n)$, and then call the set of all of these elements the real numbers. [A more precise treatment would show that the set $\mathcal{N}$ of Cauchy sequences with limit 0 forms an ideal in $\mathcal{R}$, and then that the quotient $\mathcal{R}/\mathcal{N}$ inherits the axioms for a field, and that $|\cdot|$ can be extended to $\mathcal{R}/\mathcal{N}$ with the usual properties for the real numbers satisfied with this definition of $|\cdot|$. But we are being sketchy here, and so we will not get bogged down in such details.] One can prove that $\mathbb{R}$ is complete with respect to the absolute value $|\cdot|$ inherited from $\mathbb{Q}$, and we refer to $\mathbb{R}$ as being the completion of $\mathbb{Q}$ with respect to $|\cdot|$.

We now define a substitute for the absolute value that measures the power of a given prime dividing the argument.

Definition 8.5. Let p be a prime number. Any non-zero rational number α can be written uniquely in the form $\alpha = p^r u/v$, where $u \in \mathbb{Z}$, $v \in \mathbb{N}$ and $r \in \mathbb{Z}$, such that $p \nmid uv$ and $(u, v) = 1$. We define the *p-adic valuation* $|\cdot|_p$ by setting $|0|_p = 0$, and when $\alpha \in \mathbb{Q} \setminus \{0\}$, by putting $|\alpha|_p = p^{-r}$, with r defined as above.

Exercises (i) Show that $|\alpha|_p \geqslant 0$ for all $\alpha \in \mathbb{Q}$, with equality only for $\alpha = 0$; (ii) that $|\alpha\beta|_p = |\alpha|_p|\beta|_p$ for all $\alpha, \beta \in \mathbb{Q}$; (iii) that $|\alpha + \beta|_p \leqslant \max\{|\alpha|_p, |\beta|_p\}$ for all $\alpha, \beta \in \mathbb{Q}$.

The last inequality is known as the *ultrametric inequality*, and constitutes a stronger version of the triangle inequality.

Now define Cauchy sequences in $\mathbb{Q}$ with respect to $|\cdot|_p$ just as in the classical situation above. We say that $(a_n)_{n=1}^\infty$ is Cauchy with resepct to the p-adic valuation if, whenever $\varepsilon > 0$, there exists a positive number $N(\varepsilon)$ such that whenever $n > m > N(\varepsilon)$, one has $|a_n - a_m|_p < \varepsilon$. Define next

$$\mathcal{Q}_p = \{(a_n)_{n=1}^\infty : a_n \in \mathbb{Q} \text{ for each } n, \text{ and } (a_n) \text{ is Cauchy with respect to } |\cdot|_p\}.$$

One can show that $\mathcal{Q}_p$ forms a ring under addition and multiplication defined coordinatewise in the obvious fashion. Now identify two Cauchy sequences (a_n) and (b_n) when $\lim_{n\to\infty} |a_n - b_n|_p = 0$. Modulo this equivalence, we may label Cauchy sequences, say $\alpha = (a_n)$, and then call the set of all of these elements the p-adic numbers $\mathbb{Q}_p$. [Again, a more precise treatment would show that the set $\mathcal{N}_p$ of Cauchy sequences with limit 0 forms an ideal in $\mathcal{Q}_p$, and then that the quotient $\mathcal{Q}_p/\mathcal{N}_p$ inherits the axioms for a field, and that $|\cdot|_p$ can be extended to $\mathcal{Q}_p/\mathcal{N}_p$ with properties analogous to those satisfied by $|\cdot|_p$ on $\mathbb{Q}$ enjoyed by $|\cdot|_p$ on $\mathbb{Q}_p$. Again, we are being sketchy here, and so we avoid getting bogged down in such details.] One can prove that $\mathbb{Q}_p$ is complete with respect to the p-adic valuation $|\cdot|_p$ inherited from $\mathbb{Q}$, and we refer to $\mathbb{Q}_p$ as being the completion of $\mathbb{Q}$ with respect to $|\cdot|_p$.

Example 8.6 (Conway and Sloane)**.** We give an example of a sequence in $\mathbb{Q}$ with respect to $|\cdot|_5$ that has a limit in $\mathbb{Q}_5$ that can be interpreted as $2/3$. Consider the sequence $(a_n)_{n=1}^{\infty}$ defined by $a_1 = 4$, $a_2 = 34$, $a_3 = 334$, ..., and in general $a_n = \lceil 10^n/3 \rceil$. Then for every natural number n, one has $3a_n - 2 = 10^n$, and hence $|3a_n - 2|_5 = 5^{-n}$. Thus we see that $\lim_{n\to\infty} |3a_n - 2|_5 = 0$, whence (a_n) converges in the 5-adic sense to $2/3$.

Remark 8.7. One has $\sum_{n=0}^{\infty} a_n$ converges in $\mathbb{Q}_p \iff \lim_{n\to\infty} a_n = 0$.

Write s_N for the partial sum $\sum_{n=0}^{N} a_n$. Then in order to justify this remark, note on the one hand that if $\sum_{n=0}^{\infty} a_n$ converges, then

$$\lim_{N\to\infty} a_N = \lim_{N\to\infty} (s_N - s_{N-1}) = \lim_{N\to\infty} s_N - \lim_{M\to\infty} s_M = 0.$$

On the other hand, if $\lim_{n\to\infty} a_n = 0$, then given any positive number ε, there exists a positive number $N(\varepsilon)$ such that whenever $n > N(\varepsilon)$, then one has $|a_n|_p < \varepsilon$. But then whenever $N > M > N(\varepsilon)$, one has

$$|s_N - s_M|_p = |a_{M+1} + \cdots + a_N|_p \leqslant \max_{M<n\leqslant N} |a_n|_p < \varepsilon,$$

by making use of the ultrametric inequality. Thus we see that (s_N) is a Cauchy sequence with respect to $|\cdot|_p$, and hence has a limit.

The set of p-adic numbers with valuation at most 1 is known as the p-adic *integers* $\mathbb{Z}_p$, so that $\mathbb{Z}_p = \{\alpha \in \mathbb{Q}_p : |\alpha|_p \leqslant 1\}$. Notice that the set of integers $\mathbb{Z}$ can be naturally embedded into $\mathbb{Z}_p$, and likewise $\mathbb{Q}$ can be naturally embedded into $\mathbb{Q}_p$.

Fact 8.8. If $\alpha \in \mathbb{Q}_p$, then for some non-negative integer N, one can write α in the shape

$$\alpha = \sum_{n=-N}^{\infty} a_n p^n,$$

in which the coefficients a_i lie in the set $\{0, 1, \ldots, p-1\}$.

One can check, for example, that in $\mathbb{Q}_7$, one has

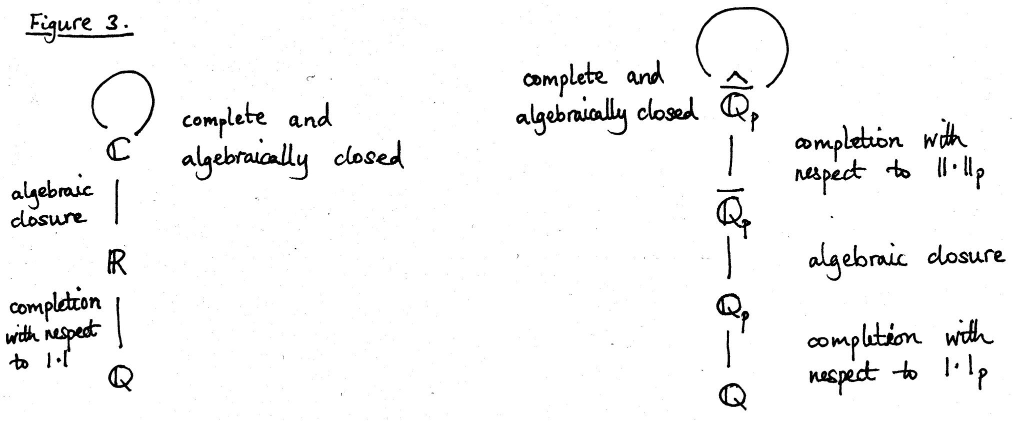
$$1/5 = 3 + 1\cdot 7 + 4\cdot 7^2 + 5\cdot 7^3 + 2\cdot 7^4 + 1\cdot 7^5 + \ldots.$$

Theorem 8.9 (Hensel's lemma revisited)**.** *Let $f \in \mathbb{Z}_p[x]$, and suppose that a is an integer satisfying the condition $|f(a)|_p < |f'(a)|_p^2$. Then there exists a unique p-adic integer α such that*

$$f(\alpha) = 0 \quad \textit{and} \quad |\alpha - a|_p \leqslant |f'(a)|_p^{-1} |f(a)|_p.$$

Example 8.10. We saw earlier that the congruence $2^2 + 1 \equiv 0 \pmod 5$ gives rise to a chain of solutions to the congruence $x^2 + 1 \equiv 0 \pmod{5^n}$. On writing $f(x) = x^2 + 1$, we have $|f(2)|_5 = |5|_5 = 5^{-1}$, and $|f'(2)|_5 = |2\cdot 2|_5 = 1$, whence $|f(2)|_5 < |f'(2)|_5^2$. Then it follows from the 5-adic version of Hensel's lemma that there exists $\alpha \in \mathbb{Z}_5$ for which $f(\alpha) = 0$ and $|\alpha - 2|_5 \leqslant 5^{-1}$. If we simply choose the truncation of the 5-adic expansion of α modulo 5^n, say α_n, then of course we obtain a solution $x = \alpha_n$ of the congruence $x^2 + 1 \pmod{5^n}$. In this sense, the 5-adic solution $x = \alpha$ of the equation $x^2 + 1 = 0$ encodes information concerning all of the associated congruences modulo 5^n.

Figure 3.


complete and algebraically closed
$\mathbb{C}$
algebraic closure
$\mathbb{R}$
completion with respect to $|\cdot|$
$\mathbb{Q}$
complete and algebraically closed
$\hat{\overline{\mathbb{Q}}}_p$
completion with respect to $\|\cdot\|_p$
$\overline{\mathbb{Q}}_p$
algebraic closure
$\mathbb{Q}_p$
completion with respect to $|\cdot|_p$
$\mathbb{Q}$

We finish this sketch of the p-adic numbers by pointing out that the interaction between completion and algebraic closure is not as simple for the p-adic numbers as for the real numbers. Thus, the completion of $\mathbb{Q}$ with respect to the ordinary absolute value $|\cdot|$ is $\mathbb{R}$, and the algebraic closure of $\mathbb{R}$ is $\mathbb{C}$, the latter being both complete and algebraically closed. Given a prime number p on the other hand, the completion of $\mathbb{Q}$ with respect to the p-adic valuation $|\cdot|_p$ is $\mathbb{Q}_p$, and the algebraic closure of $\mathbb{Q}_p$ is a larger field $\overline{\mathbb{Q}}_p$. It transpires that $\overline{\mathbb{Q}}_p$ is not itself complete (in contrast to the situation for $\mathbb{C}$). It is possible to extend the valuation $|\cdot|_p$ to a p-adic valuation $\|\cdot\|_p$ on $\overline{\mathbb{Q}}_p$, then complete the latter with respect to $\|\cdot\|_p$. The result is a field $\widehat{\overline{\mathbb{Q}}}_p$ which is both complete and algebraically closed. this represents the proper p-adic analogue of the complex numbers.

9. Primitive roots and power residues

A basic issue for understanding the structure of a multiplicative group of reduced residues is to find generators of that group, hence the idea of the order of an element and primitive roots.

Definition 9.1. Let m denote a positive integer, and let a be any integer with $(a, m) = 1$. Let h be the least positive integer with $a^h \equiv 1 \pmod{m}$. Then we say that the **order of** a **modulo** m **is** h (or that a **belongs to** h **modulo** m).

Lemma 9.2. *Let $m \in \mathbb{N}$ and $a \in \mathbb{Z}$ satisfy $(a, m) = 1$. Then the order d of a modulo m exists, and $d | \phi(m)$. Moreover, whenever $a^k \equiv 1 \pmod{m}$, one has $d|k$.*

Proof. By Euler's theorem, one has $a^{\phi(m)} \equiv 1 \pmod{m}$, and so the order of a modulo m clearly exists. Suppose then that d is the order of a modulo m, and further that $a^k \equiv 1 \pmod{m}$. Then it follows from the division algorithm that there exist integers q and r with $k = dq + r$ and $0 \leqslant r < d$. But then we obtain

$$a^k = (a^d)^q a^r \equiv a^r \equiv 1 \pmod{m},$$

whence $r = 0$. Thus we have $d|k$, and in particular we deduce that $d|\phi(m)$. □

Lemma 9.3. *Suppose that a has order h modulo m. Then a^k has order $h/(h,k)$ modulo m.*

Proof. By Lemma 9.2, one has $(a^k)^j \equiv 1 \pmod{m}$ if and only if $h|kj$. But $h|kj \iff h/(h,k)|(k/(h,k))j \iff h/(h,k)|j$. Thus the least positive integer j such that $(a^k)^j \equiv 1 \pmod{m}$ is $j = h/(h,k)$. □

Lemma 9.4. *Suppose that a has order h modulo m, and b has order k modulo m. Then whenever $(h,k) = 1$, it follows that ab has order hk modulo m.*

Proof. Let r denote the order of ab modulo m. Then since

$$(ab)^{hk} = (a^h)^k (b^k)^h \equiv 1 \pmod{m},$$

it follows from Lemma 9.2 that $r|hk$. But we also have

$$b^{rh} \equiv (a^h)^r b^{rh} \equiv (ab)^{rh} \equiv 1 \pmod{m},$$

whence $k|rh$. Since $(h,k) = 1$, moreover, the latter implies that $k|r$. Similarly, on reversing the roles of a and b, we see that $h|r$. Then since $(h,k) = 1$, we deduce that $hk|r$. We therefore conclude that $hk|r|hk$, and thus $r = hk$. □

Definition 9.5. If g belongs to the exponent $\phi(m)$ modulo m, then g is called a primitive root modulo m.

Note: If there exists a primitive root modulo m, then the multiplicative group of reduced residues modulo m is cyclic, since we have

$$(\mathbb{Z}/m\mathbb{Z})^\times = \langle g \rangle \cong C_{\phi(m)}.$$

Theorem 9.6. *If p is a prime number, then there exist $\phi(p-1)$ distinct primitive roots modulo p.*

Proof. When $p = 2$, the conclusion of the theorem is immediate, so we suppose henceforth that p is an odd prime. Observe first that each of the residues $1, 2, \ldots, p-1$ belongs to some divisor d of $p-1$ modulo p. Let $\psi(d)$ denote the number of the latter residues belonging to d modulo p. Then plainly,

$$\sum_{d|(p-1)} \psi(d) = p - 1.$$

We aim to show that for each divisor d of $p-1$, one has $\psi(d) \leqslant \phi(d)$. Given the validity of this inequality, one obtains

$$p - 1 = \sum_{d|(p-1)} \psi(d) \leqslant \sum_{d|(p-1)} \phi(d) = p - 1,$$

and so the central inequality must hold with equality for every d. The desired conclusion then follows from the case $d = p-1$ of the consequent relation $\psi(d) = \phi(d)$.

In order to verify our claim, suppose that $d|(p-1)$ and $\psi(d) \neq 0$. Let a be any residue belonging to d modulo p. It follows that $a, a^2, \ldots, a^d$ are mutually incongruent solutions of the congruence $x^d \equiv 1 \pmod{p}$. For certainly, for each positive integer j one has $(a^j)^d = (a^d)^j \equiv 1 \pmod{p}$. In addition, if it were the case that for two exponents i

and j with $1 \leqslant i < j \leqslant d$, one has $a^j \equiv a^i \pmod{p}$, then there would exist a positive integer $h = j - i < d$ with $a^h \equiv 1 \pmod{p}$, contradicting the assumption that a has order d. By Lagrange's theorem, meanwhile, there are at most d solutions modulo p to the congruence $x^d \equiv 1 \pmod{p}$, and thus the above list of residues constitutes the entire solution set modulo p. Next, on making use of Lemma 9.3, we find that whenever $(m, d) > 1$, the residue a^m has order $d/(m, d) < d$, and so the only reduced residues modulo p of order d are congruent to $a^m \pmod{p}$ for some integer m with $1 \leqslant m \leqslant d$ and $(m, d) = 1$. There are consequently precisely $\phi(d)$ such residues.

What we have shown thus far is that for each divisor d of $p - 1$, one has either $\psi(d) = \phi(d)$, or else $\psi(d) = 0$. This is a strong form of the inequality $\psi(d) \leqslant \phi(d)$ that we sought, and so our earlier discussion confirms that the number of distinct primitive roots modulo p is $\phi(p - 1)$. □

Theorem 9.7. *Suppose that g is a primitive root modulo p. Then there exists $x \in \{0, 1\}$ such that the residue $g_1 = g + px$ is a primitive root modulo p^2. When p is odd, moreover, this residue g_1 is a primitive root modulo p^k for every natural number k.*

Proof. Let g be a primitive root modulo p. Define the integer y via the relation $g^{p-1} = 1 + py$, and write $g_1 = g + px$, in which x is interpreted as a variable to be assigned in due course. In view of the expansion

$$g_1^{p-1} = (g + px)^{p-1} \equiv g^{p-1} + p(p - 1)xg^{p-2} \pmod{p^2},$$

one may write $g_1^{p-1} = 1 + pz$, in which

$$z \equiv \frac{g^{p-1} - 1}{p} + (p - 1)g^{p-2}x = y + (p - 1)g^{p-2}x \pmod{p}.$$

The coefficient of x here is not divisible by p, and so for $x = 0$ or 1 one has $(z, p) = 1$. We fix such an integer x, and now show that for every prime p this construction ensures that g_1 is a primitive root modulo p^2, and moreover that when p is odd, then the residue g_1 is a primitive root modulo p^k for every natural number k.

Suppose that g_1 has order d. Then Lemma 9.2 shows that $d|p^{k-1}(p - 1)$. But g_1 is a primitive root modulo p and $g_1^d \equiv 1 \pmod{p}$, and so in particular one has $(p - 1)|d$. Consequently, one must have $d = p^j(p - 1)$ for some integer j with $0 \leqslant j \leqslant k - 1$. But in view of our earlier observation, one has $(z, p) = 1$, and thus $g_1^{p-1} \not\equiv 1 \pmod{p^2}$. Then g_1 is always a primitive root modulo p^2. When p is odd, moreover, it follows by an inductive argument that we may write $(1 + pz)^{p^j} = 1 + p^{j+1}z_j$, for a suitable integer z_j with $(z_j, p) = 1$. Indeed, the claim holds for $j = 0$ by our construction of g_1. Moreover, if the claim holds for $0 \leqslant j \leqslant J$, then we have

$$(1 + pz)^{p^{J+1}} = (1 + p^{J+1}z_J)^p \equiv 1 + p^{J+2}z_J \pmod{p^{J+3}},$$

and thus we have $(1 + pz)^{p^{J+1}} = 1 + p^{J+2}z_{J+1}$ for some integer z_{J+1} satisfying $(z_{J+1}, p) = (z_J, p) = 1$. This confirms the inductive step. Note that this conclusion relies on the fact that $\binom{p}{2} \equiv 0 \pmod{p}$, which fails when $p = 2$. Thus we obtain the relation

$$g_1^d = (g_1^{p-1})^{p^j} = (1 + pz)^{p^j} = 1 + p^{j+1}z_j.$$

Then since g_1 has order d modulo p^k, this last expression must be congruent to 1 modulo p^k, and hence $j + 1 \geqslant k$. Then since $j \leqslant k - 1$, the only possibility is that $j = k - 1$, and we are forced to conclude that $d=\phi(p^k)$. We have shown, therefore, that g_1 is a primitive root modulo p^k, and this completes the proof of the theorem. □

Corollary 9.8. *The number of primitive roots modulo p is $\phi(p-1)$, the number modulo p^2 is $(p-1)\phi(p-1)$, and when p is odd, the number modulo p^j $(j\geqslant 3)$ is $p^{j-2}(p-1)\phi(p-1)$.*

Proof. For each modulus in question, say m, there exists a primitive root g, and moreover g^k is primitive modulo m if and only if $(k,\phi(m))=1$. But the $\phi(m)$ residues $g^k \pmod m$ are all distinct for $1\leqslant k\leqslant \phi(m)$, so every reduced residue has this form. Then the $\phi(\phi(m))$ residues $g^k \pmod m$ with $(k,\phi(m))=1$ comprise all of the primitive roots modulo m. The desired conclusion now follows on making use of the multiplicative property of the Euler totient. □

Theorem 9.9. *Let m be a natural number.*

(i) *There exists a primitive root modulo m if and only if $m=1,2,4,p^\alpha$ or $2p^\alpha$, in which p is an odd prime number and α is a natural number.*

(ii) *When $j\geqslant 3$, the order of 5 modulo 2^j is 2^{j-2}. Furthermore, every reduced residue class modulo 2^j may be written in the form $(-1)^l 5^m$, where $l\in\{0,1\}$ and $1\leqslant m\leqslant 2^{j-2}$, and in which the integers l and m are unique.*

Proof. When $m=2$, 4, the residues 1, 3, respectively, are primitive roots. When $m=p^\alpha$ the desired conclusion is immediate from Theorem 9.7. Suppose then that $m=2p^\alpha$. If g is a primitive root modulo p^α (and such exist by Theorem 9.7), then one of g and $g+p^\alpha$ is an odd integer, say g'. The order of g' modulo $2p^\alpha$ must be at least $\phi(p^\alpha)$, since g' is primitive modulo p^α. But $\phi(2p^\alpha)=\phi(2)\phi(p^\alpha)=\phi(p^\alpha)$, so that the latter observation already ensures that g' is primitive modulo $2p^\alpha$.

Suppose next that m is none of 1, 2, 4, p^α or $2p^\alpha$, for any odd prime p. Then provided that m is not a power of 2, there exist integers n_1 and n_2 with $(n_1,n_2)=1$, $n_1>n_2>2$ and $m=n_1n_2$. But then $\phi(n_1)$ and $\phi(n_2)$ are both even, whence

$$a^{\phi(m)/2}=(a^{\phi(n_1)})^{\phi(n_2)/2}\equiv 1 \pmod{n_1}\quad\text{whenever } (a,m)=1,$$

and

$$a^{\phi(m)/2}=(a^{\phi(n_2)})^{\phi(n_1)/2}\equiv 1 \pmod{n_2}\quad\text{whenever } (a,m)=1.$$

Then since $(n_1,n_2)=1$ and $m=n_1n_2$, we find that $a^{\phi(m)/2}\equiv 1 \pmod m$ whenever $(a,m)=1$. No reduced residue modulo m, therefore, has order exceeding $\phi(m)/2$, and so, in particular, no residue can be a primitive root modulo m.

It remains to consider the situation in which $m=2^j$ with $j\geqslant 3$. We begin by establishing that for each α with $\alpha\geqslant 2$, one has $2^\alpha\|(5^{2^{\alpha-2}}-1)$. This is clear when $\alpha=2$. Suppose then that the assertion holds when $\alpha=t\geqslant 2$. Then $2^t\|(5^{2^{t-2}}-1)$, whence $2\|(5^{2^{t-2}}+1)$, and thus $2^{t+1}\|(5^{2^{t-2}}-1)(5^{2^{t-2}}+1)$, or equivalently, one has $2^{t+1}\|(5^{2^{t-1}}-1)$. Then the assertion that we presently seek to establish holds with $\alpha=t+1$ whenever it holds with $\alpha=t$, whence by induction it holds for all $\alpha\geqslant 2$.

Since $2^\alpha\|(5^{2^{\alpha-2}}-1)$ for $\alpha\geqslant 2$, it follows that 5 has order precisely $2^{\alpha-2}$ modulo 2^α, and this establishes the first claim of the second part of the theorem. Observe next that there are $2^{\alpha-2}$ distinct reduced residues modulo 2^α of the shape 5^k, all of which are congruent to 1 modulo 4 (why?), and so the remaining reduced residues modulo 2^α must all be congruent to -1 modulo 4, and are hence of the shape -5^k. Thus all reduced residues modulo 2^α may be written in the form $(-1)^l 5^m$, where $l = 0$ or 1 and $1 \leqslant m \leqslant 2^{\alpha-2}$. Furthermore, these choices for l and m are distinct, for the total number of residues represented in this manner is at most $2^{\alpha-1}$, and yet there are precisely $2^{\alpha-1}$ residues to be represented. That there are no primitive roots modulo 2^α when $\alpha > 2$ follows on noting that $(-1)^l 5^m$ has order at most $2^{\alpha-2} < \phi(2^\alpha)$ when $\alpha \geqslant 3$. □

(For the cognoscenti) We have seen thus far that

$$\begin{aligned}(\mathbb{Z}/p^r\mathbb{Z})^\times &\cong C_{\phi(p^r)}, \quad \text{when } p \text{ is odd},\\ (\mathbb{Z}/2\mathbb{Z})^\times &\cong C_1,\\ (\mathbb{Z}/4\mathbb{Z})^\times &\cong C_2,\\ (\mathbb{Z}/2^r\mathbb{Z})^\times &\cong C_2 \times C_{2^{r-2}}, \quad \text{when } r \geqslant 3.\end{aligned}$$

Then on making use of the Chinese Remainder Theorem, we infer that if

$$m = 2^e \prod_{\substack{p^r \| m\\ p>2}} p^r,$$

then

$$(\mathbb{Z}/m\mathbb{Z})^\times \cong G_e \times \prod_{\substack{p^r \| m\\ p>2}} C_{\phi(p^r)},$$

where

$$G_e \cong \begin{cases} C_1, & \text{when } e = 0, 1,\\ C_2, & \text{when } e = 2,\\ C_2 \times C_{2^{e-2}}, & \text{when } e \geqslant 3.\end{cases}$$

Put

$$e(p^h) = \begin{cases} \phi(p^h), & \text{when } p \text{ is odd, and when } p^h = 2 \text{ or } 4,\\ \frac{1}{2}\phi(p^h), & \text{when } p = 2 \text{ and } h \geqslant 3,\end{cases}$$

and then define the (*Carmichael*) function

$$\lambda(n) = \operatorname*{lcm}_{p^h \| n} e(p^h).$$

It is clear from the above discussion that whenever $(a, n) = 1$, then one has

$$a^{\lambda(n)} \equiv 1 \pmod{n},$$

providing a refinement of Euler's theorem. Moroever, for every natural number n, it is apparent also that there exists an integer a with $(a, n) = 1$ having order precisely $\lambda(n)$ modulo n.

It is apparent from the above discussion that there are finite abelian groups that are not isomorphic to any multiplicative group of residues $(\mathbb{Z}/m\mathbb{Z})^\times$, for $m \in \mathbb{N}$. However, if we write the subgroup of d-th powers of $(\mathbb{Z}/m\mathbb{Z})^\times$ as

$$U_m^{(d)} = \{a^d : a \in (\mathbb{Z}/m\mathbb{Z})^\times\},$$

then it is the case that every finite abelian group is isomorphic to a group of the shape $U_m^{(d)}$ for suitable $m, d \in \mathbb{N}$. Thus, for example, one has

$$\mathbb{Z}_3 \times \mathbb{Z}_{42} \cong U_{559}^{(2)} \quad \text{and} \quad \mathbb{Z}_3 \times \mathbb{Z}_{42} \cong U_{13051}^{(10)}.$$

10. Quadratic and power residues

We now investigate residues with special properties of algebraic type.

Definition 10.1. (i) When $(a, m) = 1$ and $x^n \equiv a \pmod{m}$ has a solution, then we say that a is an **nth power residue modulo** m.
(ii) When $(a, m) = 1$, we say that a is a **quadratic residue** modulo m provided that the congruence $x^2 \equiv a \pmod{m}$ is soluble. If the latter congruence is insoluble, then we say that a is a **quadratic non-residue modulo** m.

Theorem 10.2. *Suppose that p is a prime number and $(a, p) = 1$. Then the congruence $x^n \equiv a \pmod{p}$ is soluble if and only if*

$$a^{\frac{p-1}{(n,p-1)}} \equiv 1 \pmod{p}.$$

Proof. Let g be a primitive root modulo p. Then for some natural number r one has $a \equiv g^r \pmod{p}$. If

$$a^{\frac{p-1}{(n,p-1)}} \equiv 1 \pmod{p},$$

then

$$g^{\frac{r(p-1)}{(n,p-1)}} \equiv 1 \pmod{p}.$$

But since g is primitive, the latter congruence can hold only when

$$(p-1) \,\Big|\, \frac{r(p-1)}{(n, p-1)} ,$$

whence $(n, p-1)|r$. But by the Euclidean Algorithm, there exist integers u and v with $nu + (p-1)v = (n, p-1)$, so on writing $r = k(n, p-1)$, we obtain

$$a \equiv g^{k(n,p-1)} \equiv (g^{ku})^n (g^{p-1})^{kv} \equiv (g^{ku})^n \pmod{p}.$$

Thus a is indeed an nth power residue under these circumstances.

On the other hand, if the congruence $x^n \equiv a \pmod{p}$ is soluble, then

$$a^{\frac{p-1}{(n,p-1)}} \equiv (x^{p-1})^{n/(n,p-1)} \equiv 1 \pmod{p},$$

on making use of Fermat's Little Theorem. This completes the proof of the theorem. □

Example 10.3. Determine whether or not 3 is a 4th power residue modulo 17.

Observe that on making use of Theorem 10.2, the congruence $x^4 \equiv 3 \pmod{17}$ is soluble if and only if $3^{16/4} \equiv 1 \pmod{17}$, that is, if $81 \equiv 1 \pmod{17}$. Since this congruence is not satisfied, one finds that 3 is not a 4th power residue modulo 17.

Definition 10.4. When p is an odd prime number, define the **Legendre symbol** $\left(\dfrac{a}{p}\right)$ by

$$\left(\frac{a}{p}\right) = \begin{cases} +1, & \text{when } a \text{ is a quadratic residue modulo } p, \\ -1, & \text{when } a \text{ is a quadratic non-residue modulo } p, \\ 0, & \text{when } p|a. \end{cases}$$

Theorem 10.5 (Euler's criterion)**.** *When p is an odd prime, one has*

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

Proof. If $a^{(p-1)/2} \equiv 1 \pmod{p}$, then the desired conclusion is an immediate consequence of Theorem 10.2. The conclusion is also immediate when $p|a$. It remains to consider the situation in which $a^{(p-1)/2} \not\equiv 1 \pmod{p}$. Let a be an integer with $(a,p)=1$, write $r = a^{(p-1)/2}$, and note that in view of Fermat's Little Theorem, one has $r^2 = a^{p-1} \equiv 1 \pmod{p}$, whence $r \equiv \pm 1 \pmod{p}$. Then if $r \not\equiv 1 \pmod{p}$, one necessarily has $r \equiv -1 \pmod{p}$. Thus, in the situation in which $a^{(p-1)/2} \not\equiv 1 \pmod{p}$, wherein Theorem 10.2 establishes that a is a quadratic non-residue modulo p, one has $a^{(p-1)/2} \equiv -1 \pmod{p}$, and so the desired conclusion follows once again. This completes the proof of the theorem. □

Theorem 10.6. *Let p be an odd prime number. Then*
(i) for all integers a and b, one has

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

(ii) whenever $a \equiv b \pmod{p}$, one has

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

(iii) whenever $(a,p)=1$, one has

$$\left(\frac{a^2}{p}\right) = 1 \quad \textit{and} \quad \left(\frac{a^2b}{p}\right) = \left(\frac{b}{p}\right);$$

(iv) one has

$$\left(\frac{1}{p}\right) = 1 \quad \textit{and} \quad \left(\frac{-1}{p}\right) = (-1)^{(p-1)/2}.$$

Proof. These conclusions are essentially immediate from Theorem 10.5. For example, the latter theorem shows that

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

and so the conclusion of part (i) of the theorem follows on noting that since p is odd, one cannot have $1 \equiv -1 \pmod{p}$. Parts (ii) and (iv) are trivial from the last observation, and part (iii) follows from Fermat's Little Theorem. □

Note: The number of solutions of the congruence $x^2 \equiv a \pmod{p}$ is given by $1 + \left(\frac{a}{p}\right)$. For when $(a,p)=1$ and the congruence is soluble, one has two distinct solutions and $1 + \left(\frac{a}{p}\right) = 1 + 1 = 2$. In the corresponding case in which the congruence is insoluble, one has $1 + \left(\frac{a}{p}\right) = 1 + (-1) = 0$. When $(a,p) > 1$, one the other hand, one has the single solution $x \equiv 0 \pmod{p}$, and then $1 + \left(\frac{a}{p}\right) = 1 + 0 = 1$.

The above observation provides a means of analysing the solubility of quadratic equations. For if $(a,p)=1$ and $p>3$, then the congruence $ax^2+bx+c\equiv 0 \pmod{p}$ is soluble if and only if $(2ax+b)^2\equiv b^2-4ac \pmod{p}$ is soluble, that is, if and only if either $b^2-4ac\equiv 0 \pmod{p}$, or else

$$\left(\frac{b^2-4ac}{p}\right)=1.$$

The number of solutions of the congruence is therefore precisely

$$1+\left(\frac{b^2-4ac}{p}\right).$$

For each integer a and any natural number n, define the numerically least residue of a modulo n as the integer a' satisfying $a\equiv a' \pmod{n}$ and

$$-\tfrac{1}{2}n<a'\leqslant \tfrac{1}{2}n.$$

Theorem 10.7 (Gauss' Lemma). *Let p be an odd prime number, and for each reduced residue a modulo p, let a_j denote the numerically least residue of $aj \pmod{p}$. Then*

$$\left(\frac{a}{p}\right)=(-1)^l,$$

where $l=\text{card}\{1\leqslant j\leqslant \frac{1}{2}(p-1)\,:\,a_j<0\}$.

Proof. Write $r=\frac{1}{2}(p-1)$. Then we claim that the integers $|a_j|$ $(1\leqslant j\leqslant r)$ are simply the integers $1,2,\dots,r$ in some order. In order to establish this claim, observe that for each integer j with $1\leqslant j\leqslant r$, one has $1\leqslant |a_j|\leqslant r$. Moreover, if $a_j=-a_k$ for any j and k with $1\leqslant j,k\leqslant r$, then $aj\equiv -ak \pmod{p}$, whence $a(j+k)\equiv 0 \pmod{p}$. On recalling that by hypothesis we have $(a,p)=1$, we infer that $p|(j+k)$, a conclusion that contradicts our earlier assumption that $1\leqslant j,k\leqslant r$, since then $0<j+k\leqslant 2r<p$. Thus we see that $a_j=-a_k$ for no indices j and k with $1\leqslant j,k\leqslant r$. Moreover, if $a_j=a_k$ for any j and k with $1\leqslant j,k\leqslant r$, then $aj\equiv ak \pmod{p}$, whence $j\equiv k \pmod{p}$. Our hypothesis that $1\leqslant j,k\leqslant r$ in this instance ensures that in fact $j=k$. We may therefore conclude that when $1\leqslant j,k\leqslant r$, one has $|a_j|=|a_k|$ if and only if $j=k$, and this suffices to establish our original claim.

We now complete the proof of the lemma, noting in the first instance that as an immediate consequence of the above claim, one has

$$(-1)^l r!=a_1a_2\dots a_r\equiv a(2a)\dots(ar)=a^r r! \pmod{p}.$$

Here we recall that l is the number of the reduced residues $a_1,a_2,\dots,a_r$ that have negative sign. But $p\nmid r!$, and thus we deduce that $a^r\equiv(-1)^l \pmod{p}$. The conclusion of the lemma is now immediate from Euler's criterion. □

Corollary 10.8. *When p is an odd prime number, one has*

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

Proof. When $a=2$, one has

$$a_j=\begin{cases}2j, & \text{when } 1\leqslant j\leqslant \lfloor p/4\rfloor,\\ 2j-p, & \text{when } \lfloor p/4\rfloor<j\leqslant (p-1)/2.\end{cases}$$

Then by Gauss' lemma, one has $\left(\dfrac{2}{p}\right) = (-1)^l$, where

$$l = \text{card}\{1 \leqslant j \leqslant (p-1)/2 \,:\, a_j < 0\} = \tfrac{1}{2}(p-1) - \lfloor p/4 \rfloor.$$

We now classify the odd prime numers, and find that

$$\begin{aligned}
p = 8k+1 &\Rightarrow l = \tfrac{1}{2}(8k) - \lfloor 2k + 1/4 \rfloor = 2k,\\
p = 8k-1 &\Rightarrow l = \tfrac{1}{2}(8k-2) - \lfloor 2k - 1/4 \rfloor = 2k,\\
p = 8k+3 &\Rightarrow l = \tfrac{1}{2}(8k+2) - \lfloor 2k + 3/4 \rfloor = 2k+1,\\
p = 8k-3 &\Rightarrow l = \tfrac{1}{2}(8k-4) - \lfloor 2k - 3/4 \rfloor = 2k-1.
\end{aligned}$$

Thus, on noting that

$$((8k \pm 1)^2 - 1)/8 \equiv 0 \pmod 2 \quad \text{and} \quad ((8k \pm 3)^2 - 1)/8 \equiv 1 \pmod 2,$$

we find that $l \equiv (p^2-1)/8 \pmod 2$, and thus the conclusion of the corollary follows from Gauss' lemma. □

Note also that from Euler's criterion, one has

$$\left(\frac{1}{p}\right) = 1 \quad \text{and} \quad \left(\frac{-1}{p}\right) = (-1)^{(p-1)/2}.$$

Thus we have simple formulae for $\left(\dfrac{\pm 1}{p}\right)$ and $\left(\dfrac{\pm 2}{p}\right)$, and it is clear from the multiplicative property of $\left(\dfrac{\cdot}{p}\right)$ that it suffices now to compute $\left(\dfrac{q}{p}\right)$ for odd prime numbers q in order to calculate $\left(\dfrac{a}{p}\right)$ in general.

11. The law of quadratic reciprocity

Theorem 11.1 (Gauss)**.** *Let p and q be distinct odd prime numbers. Then*

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{(p-1)(q-1)/4}.$$

Note 11.2. *Rewriting the expression on the right hand side of the last equation in the shape*

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

we see that $\left(\dfrac{p}{q}\right) = \left(\dfrac{q}{p}\right)$ unless p and q are **both** *congruent to* 3 *modulo* 4.

Proof. (of the law of quadratic reciprocity) Observe that, as a consequence of Gauss' lemma, one has that $\left(\dfrac{p}{q}\right) = (-1)^l$, where l is the number of lattice points (x, y) satisfying the inequalities

$$0 < x < q/2 \quad \text{and} \quad -q/2 < px - qy < 0.$$

But y is an integer, and

$$y < \frac{px}{q} + \frac{1}{2} < (p+1)/2.$$

Thus l is the number of lattice points (x,y) in the rectangle $\mathcal{R}$ defined by $0 < x < q/2$ and $0 < y < p/2$ which satisfy $-q/2 < px - qy < 0$ (see Fig. 4). Similarly, we have $\left(\dfrac{q}{p}\right) = (-1)^m$, where m is the number of lattice points in the same rectangle $\mathcal{R}$ with $-p/2 < qy - px < 0$ (see Fig. 5).

Figure 4.

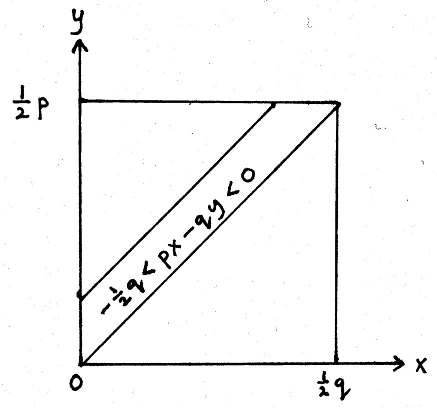
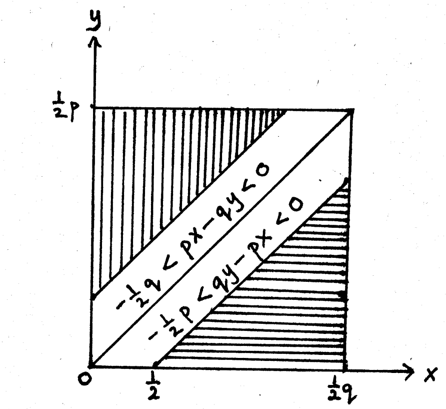


Figure 5.



We therefore obtain

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

and this will yield the desired conclusion

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{(p-1)(q-1)/4},$$

provided that $\frac{1}{2}(p-1)\cdot\frac{1}{2}(q-1) - (l+m)$ is even. But the latter quantity is simply the number of lattice points (x,y) contained in the shaded region in Fig. 5, namely those lattice points satisfying

$$px - qy \leqslant -q/2 \quad \text{or} \quad qy - px \leqslant -p/2.$$

These two regions are disjoint, and contain the same number of points, as can be seen by considering the bijective correspondence

$$(x,y) \longleftrightarrow (\tfrac{1}{2}(q+1) - x, \tfrac{1}{2}(p+1) - y).$$

Note here that $\frac{1}{2}q + px - qy$ is in bijective correspondence with

$$\tfrac{1}{2}q + p\left(\tfrac{1}{2}(q+1) - x\right) - q\left(\tfrac{1}{2}(p+1) - y\right) = \tfrac{1}{2}p + qy - px,$$

and that the ordered pair $\left(\frac{1}{2}(q+1) - x, y\right)$ is in bijective correspondence with $(x, \frac{1}{2}(p+1) - y)$. Moreover, $x = 1$ is mapped to $x = \frac{1}{2}(q-1)$ and likewise $y = 1$ to $y = \frac{1}{2}(p-1)$, and vice versa. Thus the number of lattice points in the shaded region of Fig. 5 is twice the number contained in either shaded triangle, and hence is even. This proves that

$\frac{1}{2}(p-1)\frac{1}{2}(q-1)-(l+m)$ is even, as we sought to establish, and hence the conclusion of the theorem follows as described above. □

Example 11.3. Determine the value of $\left(\dfrac{-3}{p}\right)$.

By Quadratic Reciprocity we have

$$\left(\frac{3}{p}\right)\left(\frac{p}{3}\right)=(-1)^{(3-1)(p-1)/4}=(-1)^{(p-1)/2},$$

and by Euler's criterion, on the other hand,

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

Thus we see that

$$\left(\frac{-3}{p}\right)=\left(\frac{-1}{p}\right)\left(\frac{3}{p}\right)=(-1)^{(p-1)/2}\cdot(-1)^{(p-1)/2}\left(\frac{p}{3}\right)=\left(\frac{p}{3}\right).$$

But

$$\left(\frac{p}{3}\right)=\begin{cases}\left(\dfrac{1}{3}\right)=1, & \text{when } p\equiv 1 \pmod 3,\\ \left(\dfrac{2}{3}\right)=-1, & \text{when } p\equiv 2 \pmod 3.\end{cases}$$

Thus we deduce that

$$\left(\frac{-3}{p}\right)=\begin{cases}1, & \text{when } p\equiv 1 \pmod 3,\\ -1, & \text{when } p\equiv 2 \pmod 3.\end{cases}$$

One can use this evaluation to show that the only possible prime divisors of x^2+3, for integral values of x, are 3 and primes p with $p\equiv 1 \pmod 3$. From here, an argument similar to that due to Euclid shows that there are infinitely many primes congruent to 1 modulo 3.

Example 11.4. Determine the value of $\left(\dfrac{21}{71}\right)$.

Applying the multiplicative property of the Legendre symbol, followed by quadratic reciprocity, one finds that

$$\begin{aligned}\left(\frac{21}{71}\right)&=\left(\frac{3}{71}\right)\left(\frac{7}{71}\right)=(-1)^{(71-1)(3-1)/4+(71-1)(7-1)/4}\left(\frac{71}{3}\right)\left(\frac{71}{7}\right)\\&=\left(\frac{71}{3}\right)\left(\frac{71}{7}\right)=\left(\frac{2}{3}\right)\left(\frac{1}{7}\right)=\left(\frac{2}{3}\right)=-1.\end{aligned}$$

So $\left(\dfrac{21}{71}\right)=-1$, and hence 21 is not a quadratic residue modulo 71.

12. The Jacobi symbol

We wish to generalise the Legendre symbol $\left(\frac{\cdot}{p}\right)$ to accomodate composite moduli.

Definition 12.1. Let Q be a positive odd integer, and suppose that $Q = p_1 \dots p_s$, where the p_i are prime numbers (not necessarily distinct). Then we define the **Jacobi symbol** $\left(\frac{a}{Q}\right)$ as follows:

(i) $\left(\frac{a}{1}\right) = 1$;

(ii) $\left(\frac{a}{Q}\right) = 0$ whenever $(a, Q) > 1$;

(iii) $\left(\frac{a}{Q}\right) = \left(\frac{a}{p_1}\right)\left(\frac{a}{p_2}\right)\dots\left(\frac{a}{p_s}\right)$ whenever $(a, Q) = 1$.

Just as in the discussion concerning the Legendre symbol, we begin with some simple properties of the Jacobi symbol.

Theorem 12.2. *Suppose that Q and Q' are positive odd integers. Then:*

(i) $\left(\frac{P}{Q}\right)\left(\frac{P}{Q'}\right) = \left(\frac{P}{QQ'}\right)$*;*

(ii) $\left(\frac{P}{Q}\right)\left(\frac{P'}{Q}\right) = \left(\frac{PP'}{Q}\right)$*;*

(iii) whenever $(P, Q) = 1$, one has $\left(\frac{P}{Q^2}\right) = \left(\frac{P^2}{Q}\right) = 1$*;*

(iv) whenever $(PP', QQ') = 1$, one has $\left(\frac{P'P^2}{Q'Q^2}\right) = \left(\frac{P'}{Q'}\right)$*;*

(v) whenever $P \equiv P' \pmod{Q}$, one has $\left(\frac{P}{Q}\right) = \left(\frac{P'}{Q}\right)$*.*

Proof. Part (i) is immediate from the definition of the Jacobi symbol, and part (ii) is immediate from the properties of the Legendre symbol. Parts (iii) and (iv) follow directly from parts (i) and (ii), since the Jacobi symbol takes values 0 or ±1. For part (v) of the theorem, observe that whenever $P \equiv P' \pmod{Q}$, one has $P \equiv P' \pmod{p}$ for each prime number p dividing Q, whence also $\left(\frac{P}{p}\right) = \left(\frac{P'}{p}\right)$ for each prime p dividing Q. The desired conclusion is therefore again immediate from the definition of the Jacobi symbol. □

Note 12.3. *If the Jacobi symbol $\left(\frac{a}{n}\right) = -1$, then it follows that a is* **not** *a quadratic residue modulo n, since for some prime p with $p|n$ one must have that the Legendre symbol $\left(\frac{a}{p}\right) = -1$.* **But** *if $\left(\frac{a}{n}\right) = 1$, then it is* **not** *necessarily the case that a is a quadratic residue modulo n. For example, one has*

$$\left(\frac{2}{15}\right) = 1, \quad \textit{but} \quad \left(\frac{2}{3}\right) = -1 \quad \textit{and} \quad \left(\frac{2}{5}\right) = -1.$$

The Jacobi symbol remains useful for calculating Legendre symbols, because it satisfies the same reciprocity and simplifying relations as the Legendre symbol (as we now demonstrate), and at the same time, whenever the Legendre symbol $\left(\dfrac{a}{n}\right)$ is defined (that is, provided that n is an odd prime number), then its value is the same as that of the corresponding Jacobi symbol.

Theorem 12.4. *Suppose that Q is a positive odd integer. Then*

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

Proof. Suppose that Q is odd, and that $Q = p_1 \dots p_s$ with each p_i a prime number. Then

$$\left(\frac{-1}{Q}\right) = \prod_{i=1}^{s} \left(\frac{-1}{p_i}\right) = \prod_{i=1}^{s} (-1)^{(p_i-1)/2}.$$

But whenever n_1 and n_2 are both odd, one has $\frac{1}{2}(n_1-1)(n_2-1) \equiv 0 \pmod 2$, whence

$$\tfrac{1}{2}(n_1-1) + \tfrac{1}{2}(n_2-1) \equiv \tfrac{1}{2}(n_1n_2-1) - \tfrac{1}{2}(n_1-1)(n_2-1) \equiv \tfrac{1}{2}(n_1n_2-1) \pmod 2.$$

Iterating the latter relation, we deduce that

$$\tfrac{1}{2}(Q-1) \equiv \sum_{i=1}^{s} (p_i-1)/2 \pmod 2,$$

whence $\left(\dfrac{-1}{Q}\right) = (-1)^{(Q-1)/2}$.

Similarly, we have

$$\left(\frac{2}{Q}\right) = \prod_{i=1}^{s} \left(\frac{2}{p_i}\right) = \prod_{i=1}^{s} (-1)^{(p_i^2-1)/8}.$$

But whenever n_1 and n_2 are both odd, it follows that

$$\tfrac{1}{8}(n_1^2-1)(n_2^2-1) \equiv 0 \pmod 2,$$

whence

$$\begin{aligned}\tfrac{1}{8}(n_1^2-1) + \tfrac{1}{8}(n_2^2-1) &= \tfrac{1}{8}(n_1^2n_2^2-1) - \tfrac{1}{8}(n_1^2-1)(n_2^2-1)\\ &\equiv \tfrac{1}{8}(n_1^2n_2^2-1) \pmod 2.\end{aligned}$$

Thus, again iterating this relation, we find that

$$(Q^2-1)/8 \equiv \sum_{i=1}^{s} (p_i^2-1)/8 \pmod 2,$$

whence

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

□

Theorem 12.5 (Quadratic Reciprocity)**.** *Suppose that P and Q are odd positive integers with $(P,Q)=1$. Then*

$$\left(\frac{P}{Q}\right)\left(\frac{Q}{P}\right) = (-1)^{(P-1)(Q-1)/4}.$$

Proof. Suppose that $Q = q_1 \dots q_s$ and $P = p_1 \dots p_r$ are factorisations of P and Q, respectively, into products of prime numbers. Then we have

$$\left(\frac{P}{Q}\right) = \prod_{j=1}^{s} \left(\frac{P}{q_j}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s} \left(\frac{p_i}{q_j}\right).$$

Then by quadratic reciprocity for the Legendre symbol, we obtain

$$\left(\frac{P}{Q}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s} (-1)^{(p_i-1)(q_j-1)/4} \left(\frac{q_j}{p_i}\right) = (-1)^{\omega} \left(\frac{Q}{P}\right),$$

where we write

$$\omega = \sum_{i=1}^{r}\sum_{j=1}^{s} (p_i-1)(q_j-1)/4.$$

But as in the proof of Theorem 12.4, one has

$$\begin{aligned}\sum_{i=1}^{r}\sum_{j=1}^{s} (p_i-1)(q_j-1)/4 &= \left(\sum_{i=1}^{r}(p_i-1)/2\right)\left(\sum_{j=1}^{s}(q_j-1)/2\right) \\ &\equiv \tfrac{1}{2}(P-1)\cdot\tfrac{1}{2}(Q-1) \pmod{2}.\end{aligned}$$

We therefore deduce that

$$\left(\frac{P}{Q}\right) = (-1)^{(P-1)(Q-1)/4} \left(\frac{Q}{P}\right),$$

and the conclusion of the theorem now follows immediately. □

Jacobi symbols are useful for calculating Legendre symbols, since they take the same values for prime moduli, and one can skip intermediate factorisations before applying reciprocity.

Example 12.6. Calculate the Legendre symbol $\left(\frac{1111}{8093}\right)$.

One has

$$\begin{aligned}\left(\frac{1111}{8093}\right) &= (-1)^{(8092)(1110)/4}\left(\frac{8093}{1111}\right) = \left(\frac{316}{1111}\right) = \left(\frac{2}{1111}\right)^2\left(\frac{79}{1111}\right) \\ &= (-1)^{(1110)(78)/4}\left(\frac{1111}{79}\right) = -\left(\frac{5}{79}\right) = -(-1)^{(4)(78)/4}\left(\frac{79}{5}\right) \\ &= -\left(\frac{4}{5}\right) = -\left(\frac{2}{5}\right)^2 = -1.\end{aligned}$$

So 1111 is not a quadratic residue modulo 8093.

Example 12.7. Determine whether or not the congruence $x^2 + 6x - 50 \equiv 0 \pmod{79}$ has a solution.

Observe that $x^2+6x-50=(x+3)^2-59$, and hence $x^2+6x-50\equiv 0 \pmod{79}$ has a solution if and only if $\left(\frac{59}{79}\right)=1$. But

$$\begin{aligned}\left(\frac{59}{79}\right)&=\left(\frac{-20}{79}\right)=\left(\frac{-1}{79}\right)\left(\frac{2}{79}\right)^2\left(\frac{5}{79}\right)=(-1)^{(79-1)/2}\left(\frac{5}{79}\right)\\&=-(-1)^{(5-1)(79-1)/4}\left(\frac{79}{5}\right)=-\left(\frac{4}{5}\right)=-1.\end{aligned}$$

Hence the congruence $x^2+6x-50\equiv 0 \pmod{79}$ has no solution.

Example 12.8. Find the number of solutions of the congruence

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

when p is an odd prime.

Observe first that this number N is equal to

$$\begin{aligned}\sum_{x=1}^{p}\left(1+\left(\frac{x^2+1}{p}\right)\right)&=p+\sum_{y=1}^{p}\left(1+\left(\frac{y}{p}\right)\right)\left(\frac{y+1}{p}\right)\\&=p+\sum_{y=1}^{p}\left(\frac{y+1}{p}\right)+\sum_{y=1}^{p}\left(\frac{y(y+1)}{p}\right).\end{aligned}$$

Next we note that

$$\sum_{y=1}^{p}\left(\frac{y+1}{p}\right)=\sum_{v=1}^{p}\left(\frac{v}{p}\right)=0.$$

For if g is a primitive root modulo p, then $\left(\frac{g}{p}\right)=-1$ (why?), and hence

$$\sum_{v=1}^{p}\left(\frac{v}{p}\right)=\sum_{v=1}^{p-1}\left(\frac{v}{p}\right)=\sum_{l=1}^{p-1}\left(\frac{g^l}{p}\right)=\sum_{l=1}^{p-1}\left(\frac{g}{p}\right)^l=\sum_{l=1}^{p-1}(-1)^l=0.$$

We observe next that since for $(x,p)=1$ one has

$$\left(\frac{x^{-1}}{p}\right)\left(\frac{x}{p}\right)=\left(\frac{x^{-1}x}{p}\right)=\left(\frac{1}{p}\right)=1,$$

then

$$\left(\frac{x}{p}\right)=\left(\frac{x^{-1}}{p}\right).$$

Hence

$$\begin{aligned}\sum_{x=1}^{p}\left(\frac{x(x+1)}{p}\right)&=\sum_{x=1}^{p-1}\left(\frac{x(x+1)}{p}\right)=\sum_{x=1}^{p-1}\left(\frac{x}{p}\right)\left(\frac{x+1}{p}\right)\\&=\sum_{x=1}^{p-1}\left(\frac{x^{-1}}{p}\right)\left(\frac{x+1}{p}\right)=\sum_{x=1}^{p-1}\left(\frac{x^{-1}(x+1)}{p}\right)\\&=\sum_{x=1}^{p-1}\left(\frac{1+x^{-1}}{p}\right)=\sum_{y=1}^{p-1}\left(\frac{1+y}{p}\right)\\&=\sum_{y=1}^{p}\left(\frac{1+y}{p}\right)-\left(\frac{1}{p}\right)=\sum_{z=1}^{p}\left(\frac{z}{p}\right)-1=-1.\end{aligned}$$

Thus

$$\sum_{x=1}^{p}\left(\frac{x(x+1)}{p}\right)=-1.$$

We therefore conclude that

$$N=p-1.$$

13. Application of residue symbols to counting solutions of congruences

Let p be an odd prime. One can verify that the sum over y of the quadratic residue symbol $\left(\frac{y}{p}\right)$ is zero, either by making use of the conclusions of Question 1 on Problem Sheet 7 (which is short and direct), or as follows (which is more obscure). Write

$$M=\sum_{y=1}^{p}\left(\frac{y}{p}\right)=\sum_{y=1}^{p-1}\left(\frac{y}{p}\right).$$

When $(a,p)=1$, the mapping $y\mapsto ay \pmod p$ permutes the reduced residues modulo p. Moreover, a primitive root g modulo p must be a quadratic non-residue, in view of Euler's criterion (we have $g^{(p-1)/2}\not\equiv 1 \pmod p$). Then we have

$$M=\sum_{y=1}^{p-1}\left(\frac{gy}{p}\right)=\left(\frac{g}{p}\right)\sum_{y=1}^{p-1}\left(\frac{y}{p}\right)=\left(\frac{g}{p}\right)M.$$

But $\left(\frac{g}{p}\right)=-1$, and so we are forced to conlcude that $M=0$.

Theorem 13.1. *Let $f(x)=ax^2+bx+c$, where a, b and c are integers, and let p be an odd prime number. Suppose that $(a,p)=1$, and write $d=b^2-4ac$. Then if $p\nmid d$, one has*

$$\sum_{x=1}^{p}\left(\frac{f(x)}{p}\right)=-\left(\frac{a}{p}\right),$$

and if $p|d$, then

$$\sum_{x=1}^{p}\left(\frac{f(x)}{p}\right)=(p-1)\left(\frac{a}{p}\right).$$

Proof. One has

$$4a(ax^2+bx+c)=(2ax+b)^2-(b^2-4ac)=y^2-d,$$

say, where $d=b^2-4ac$ and $y=2ax+b$. Then if $p|d$, we obtain

$$\sum_{x=1}^{p}\left(\frac{f(x)}{p}\right)=\sum_{y=1}^{p}\left(\frac{(4a)^{-1}}{p}\right)\left(\frac{y^2}{p}\right)=\sum_{y=1}^{p-1}\left(\frac{4a}{p}\right)=(p-1)\left(\frac{a}{p}\right).$$

This establishes the second claim in the statement of the theorem.

Suppose then that $p\nmid d$. Then

$$\left(\frac{4a}{p}\right)\sum_{x=1}^{p}\left(\frac{f(x)}{p}\right)=\sum_{x=1}^{p}\left(\frac{(2ax+b)^2-(b^2-4ac)}{p}\right)=\sum_{y=1}^{p}\left(\frac{y^2-d}{p}\right).$$

Write

$$S(d)=\sum_{y=1}^{p}\left(\frac{y^2-d}{p}\right).$$

Then we see that

$$\sum_{x=1}^{p}\left(\frac{f(x)}{p}\right)=\left(\frac{4a}{p}\right)S(d)=\left(\frac{a}{p}\right)S(d).$$

But since $1+\left(\frac{z}{p}\right)$ is non-zero only when z is a square modulo p, say y^2, and is 2 when the latter is non-zero, and 1 when zero, we find that

$$\begin{aligned}S(d)=\sum_{y=1}^{p}\left(\frac{y^2-d}{p}\right)&=\sum_{z=1}^{p}\left(\left(\frac{z}{p}\right)+1\right)\left(\frac{z-d}{p}\right)\\&=\sum_{z=1}^{p}\left(\frac{z(z-d)}{p}\right)+\sum_{z=1}^{p}\left(\frac{z-d}{p}\right).\end{aligned}$$

The last sum is zero, in view of the comments in the preamble to the statement of Theorem 13.1. Thus, on making the change of variable $z=wd$, we find that

$$S(d)=\sum_{w=1}^{p}\left(\frac{dw(dw-d)}{p}\right)=\sum_{w=1}^{p}\left(\frac{d}{p}\right)^2\left(\frac{w(w-1)}{p}\right)=\sum_{w=1}^{p}\left(\frac{w(w-1)}{p}\right).$$

Thus $S(d)$ is independent of d, say $S(d)=S(1)$. But then we deduce that

$$(p-1)S(1)=\sum_{d=1}^{p-1}S(d)=\sum_{d=1}^{p-1}\sum_{y=1}^{p}\left(\frac{y^2-d}{p}\right)=\sum_{d=1}^{p}\sum_{y=1}^{p}\left(\frac{y^2-d}{p}\right)-\sum_{y=1}^{p}\left(\frac{y^2}{p}\right).$$

We now make the change of variable $u = y^2 - d$ in the first summation of the penultimate term, and again make use of the comments at the outset of this section, and thereby deduce that

$$(p-1)S(1) = \sum_{y=1}^{p} \sum_{u=1}^{p} \left(\frac{u}{p}\right) - (p-1) = -(p-1).$$

Then $S(1) = -1$, whence $S(d) = -1$ for all d with $(d,p) = 1$. We may therefore conclude that when $p \nmid d$, one has

$$\sum_{x=1}^{p} \left(\frac{f(x)}{p}\right) = -\left(\frac{a}{p}\right).$$

This completes the proof of the theorem. □

When p is an odd prime with $p \nmid a$, the conclusion of Theorem 13.1 shows that the number of solutions of the congruence $y^2 \equiv ax^2 + bx + c \pmod{p}$ with $1 \leqslant x, y \leqslant p$ is

$$\sum_{x=1}^{p} \left(1 + \left(\frac{ax^2+bx+c}{p}\right)\right) = p - \left(\frac{a}{p}\right),$$

when $p \nmid (b^2 - 4ac)$, and

$$\sum_{x=1}^{p} \left(1 + \left(\frac{ax^2+bx+c}{p}\right)\right) = p + (p-1)\left(\frac{a}{p}\right),$$

when $p | (b^2 - 4ac)$. In both situations, the number of solutions of the congruence is positive, and so the congruence $y^2 \equiv ax^2 + bx + c \pmod{p}$ is soluble.

14. Arithmetical functions

Recall from the preamble to the statement of Theorem 5.5 that a function $f : \mathbb{N} \to \mathbb{C}$ is called an arithmetical function. Recall also that a multiplicative function f satisfies the property that whenever $(m,n) = 1$, one has $f(mn) = f(m)f(n)$. Also (Lemma 5.6), the function $g(n) = \sum_{d|n} f(d)$ is multiplicative whenever $f(n)$ is multiplicative. In this section we discuss various properties of arithmetical functions, many of them multiplicative, and seek to understand what they "look" like.

Examples of arithmetical functions:
(i) the *divisor function* $d(n)$, or $\tau(n)$, is defined for $n \in \mathbb{N}$ by $\tau(n) = \sum_{d|n} 1$. The divisor function is therefore multiplicative, as a consequence of Lemma 5.6. The *k-fold divisor function* $d_k(n)$, or $\tau_k(n)$, is defined via the relation

$$\tau_k(n) = \sum_{\substack{d_1 \dots d_k = n \\ d_1, \dots, d_k \in \mathbb{N}}} 1 \quad \text{or} \quad \tau_k(n) = \sum_{d|n} \tau_{k-1}(d).$$

This function is also multiplicative, as a consequence again of Lemma 5.6.
(ii) the *sum of divisors function* $\sigma(n)$ is defined by $\sigma(n) = \sum_{d|n} d$, and so is multiplicative by Lemma 5.6. Similarly, the sum of kth powers of divisors function $\sigma_k(n) = \sum_{d|n} d^k$ is also multiplicative.

(iii) the *number of distinct prime divisors function* $\omega(n)$ is defined by

$$\omega(n) = \sum_{\substack{p|n \\ p \text{ prime}}} 1,$$

so that if $n = \prod_{i=1}^t p_i^{r_i}$ is the canonical prime factorisation of n, then $\omega(n) = t$. Note that for any non-zero complex number c, the function $c^{\omega(n)}$ is multiplicative (easy to check this!).

(iv) the *number of prime divisors function* $\Omega(n)$ is defined by

$$\Omega(n) = \sum_{\substack{p^r \| n \\ p \text{ prime}}} r,$$

so that if $n = \prod_{i=1}^t p_i^{r_i}$ is the canonical prime factorisation of n, then $\Omega(n) = r_1 + r_2 + \cdots + r_t$.

(v) the Euler totient $\phi(n)$ is a multiplicative function (Theorem 5.5).

(vi) the Möbius function $\mu(n)$ is defined for natural numbers n by

$$\mu(n) = \begin{cases} (-1)^{\omega(n)}, & \text{when } n \text{ is squarefree,} \\ 0, & \text{otherwise.} \end{cases}$$

Here, by a *squarefree number*, we mean an integer that is not divisible by the square of any prime number. Thus, if $n = p_1 p_2 \dots p_k$ with $p_1, \dots, p_k$ distinct primes, one has $\mu(n) = (-1)^k$, and it follows easily that $\mu(n)$ is multiplicative.

Note that, just as with our earlier discussion of the Euler totient, a function that is multiplicative will be relatively easy to evaluate when its argument has a known prime factorisation. For example, one can see rather easily that when p is a prime number and h is a non-negative integer, then $\tau(p^h) = h + 1$, and

$$\sigma(p^h) = \sum_{r=0}^{h} p^r = \frac{p^{r+1} - 1}{p - 1},$$

and thus

$$\tau(n) = \prod_{p^r \| n} (r+1) \quad \text{and} \quad \sigma(n) = \prod_{p^r \| n} \left(\frac{p^{r+1} - 1}{p - 1} \right).$$

The Möbius inversion formulae The arithmetic function defined by (vi) above, the Möbius function, has special properties that make it particularly useful in studying averages of other arithmetic functions (and much else besides). Recall that

$$\mu(n) = \begin{cases} (-1)^{\omega(n)}, & \text{when } n \text{ is squarefree,} \\ 0, & \text{otherwise.} \end{cases}$$

We define a rather trivial multiplicative function $\nu(n)$ by

$$\nu(n) = \begin{cases} 1, & \text{when } n = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 14.1. *One has* $\sum_{d|n} \mu(d) = \nu(n)$.

Proof. Since $\mu(n)$ is a multiplicative function of n, it follows from Lemma 5.6 that $\sum_{d|n}\mu(d)$ is also multiplicative. But on writing $f(n)=\sum_{d|n}\mu(d)$, one finds that

$$f(p^\alpha)=\sum_{h=0}^{\alpha}\mu(p^h)=1-1=0,\quad \text{for } \alpha>0,$$

and $f(1)=\mu(1)=1$. Thus, in view of the multiplicativity of $f(n)$, one finds that $f(n)$ is zero unless n has no prime divisors, a circumstance that occurs only when $n=1$. This completes the proof of the theorem. □

We can now describe a certain duality between arithmetic functions, and functions defined via divisor sums.

Theorem 14.2 (the Möbius inversion formulae)**.** *(i) Let f be any arithmetical function, and define*

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

Then one has

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

(ii) Suppose that g is any arithmetical function, and define

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

Then one has

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

Proof. (i) Given that $g(n)=\sum_{d|n}f(d)$, one obtains

$$\begin{aligned}\sum_{d|n}\mu(d)g(n/d)&=\sum_{d|n}\sum_{e|(n/d)}\mu(d)f(e)=\sum_{e|n}f(e)\sum_{d|(n/e)}\mu(d)\\&=\sum_{e|n}f(e)\nu(n/e)=f(n).\end{aligned}$$

(ii) Given that $f(n)=\sum_{d|n}\mu(d)g(n/d)$, one obtains

$$\begin{aligned}\sum_{d|n}f(d)=\sum_{d|n}f(n/d)&=\sum_{d|n}\sum_{e|(n/d)}\mu(e)g(n/(de))\\&=\sum_{e|n}\sum_{d|(n/e)}\mu(e)g(n/(de))=\sum_{e|n}\sum_{d|(n/e)}\mu(e)g(d)\\&=\sum_{d|n}g(d)\sum_{e|(n/d)}\mu(e)=\sum_{d|n}g(d)\nu(n/d)=g(n).\end{aligned}$$

□

Note that Möbius inversion applies to all arithmetical functions, without any hypothesis concerning whether or not they are multiplicative.

Example 14.3. Recall that we showed in the corollary to Lemma 5.6 that $\sum_{d|n} \phi(d) = n$. As an immediate consequence of the Möbius inversion formulae, we deduce that

$$\phi(n) = \sum_{d|n} \mu(d)(n/d) = n \sum_{d|n} \mu(d)/d.$$

Perfect numbers.

Definition 14.4 (Perfect numbers)**.** A natural number n, for which the sum of the positive divisors smaller than n is equal to n, is called a **perfect number**.

Equivalently, the natural number n is perfect if and only if $\sigma(n) = 2n$. Examples include $6 = 1 + 2 + 3$, $28 = 1 + 2 + 4 + 7 + 14$, and 496.

Theorem 14.5. *The natural number n is an even perfect number if and only if $n = 2^{p-1}(2^p - 1)$ for some prime number p for which $2^p - 1$ is prime.*

Proof. In one direction, observe that under the hypotheses of the statement of the theorem, one has

$$\begin{aligned}\sigma(2^{p-1}(2^p - 1)) &= \sum_{l=0}^{p-1} 2^l(2^p - 1) + \sum_{m=0}^{p-1} 2^m \\ &= (2^p - 1)^2 + (2^p - 1) = 2\left(2^{p-1}(2^p - 1)\right),\end{aligned}$$

so that $2^{p-1}(2^p - 1)$ is indeed perfect.

In the other direction, suppose that n is an even perfect number, say $n = 2^k m$, with m an odd number. Then since n is perfect, one has $\sigma(n) = 2n$, and so from the multiplicative property of $\sigma(\cdot)$, one has

$$2^{k+1}m = 2(2^k m) = \sigma(2^k m) = \sigma(2^k)\sigma(m) = (2^{k+1} - 1)\sigma(m),$$

whence $2^{k+1}|\sigma(m)$, say $\sigma(m) = 2^{k+1}l$. On substituting back into the previous relation, we deduce that $m = (2^{k+1} - 1)l$. If $l > 1$, then m has distinct divisors 1, m and l, whence $\sigma(m) \geqslant l + m + 1 > l + m = 2^{k+1}l = \sigma(m)$, which yields a contradiction. We are therefore forced to conclude that $l = 1$, whence $\sigma(m) = m + 1$, and so m is prime. Thus we find that $n = (2^{k+1} - 1)2^k$ with $2^{k+1} - 1$ prime, and the latter implies that $k + 1$ is itself prime (convince youself as to why this is true!). □

Notice that Theorem 14.5 shows that any even perfect number has the form

$$\frac{1}{2} M_p(M_p + 1),$$

where M_p is the Mersenne prime $2^p - 1$. It is suspected that there are infinitely many Mersenne primes, though this remains unproved, and hence infinitely many even perfect numbers. For more on this subject, see GIMPS, the Great Internet Mersenne Prime Search, at http://www.mersenne.org/, for more on Mersenne primes. The largest known perfect number is $2^{136279840}(2^{136279841} - 1)$ (see p.2). No odd perfect numbers are known, and it is conjectured that there are none. If an odd perfect number exists, then it has at least nine distinct prime factors (see P. P. Nielsen, *Odd perfect numbers have at least nine distinct prime factors*, Math. Comp. 76 (2007), 2109-2126) and at least 300 digits in base ten (see R. P. Brent, G. L. Cohen and H. J. J. te Riele, *Improved techniques for lower bounds for odd perfect numbers*, Math. Comp. 57 (1991), 857-868).

15. Estimates for arithmetical functions

We now explore the "population statistics" of values of arithmetical functions: what is the maximal/minimal size of such a function, the average size, the variance, etc? In order properly to discuss such issues, we need to recall some standard analytic notation.

Given functions $f, g : \mathbb{R} \to \mathbb{R}$, with g taking positive values, we write $f(x) = O(g(x))$ (for $x \geqslant x_0$) when there exists some positive constant C for which $|f(x)| \leqslant Cg(x)$ (for $x \geqslant x_0$).

Example 15.1. One has $x = O(x^2)$ for $x \geqslant 1$, $1/x^2 = O(1)$ for $x \geqslant 1$, and $x = O(e^x)$ for $x \geqslant 0$.

There are two useful strategies to keep in mind when addressing questions concerning estimates for arithmetic functions:
(i) in order to estimate the *size* of a *multiplicative* function $f(n)$, one should first estimate $f(\cdot)$ on prime powers, and then combine this information with knowledge about the distribution of prime numbers;
(ii) If one wishes to estimate the average size of an arithmetical function $g(n)$, one can apply the Möbius inversion formulae to write $g(n)$ in the shape

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

in which

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

Frequently, one finds that this new function $f(n)$ is reasonably well-behaved, and then one has

$$\sum_{1 \leqslant n \leqslant x} g(n) = \sum_{1 \leqslant n \leqslant x} \sum_{d|n} f(d) = \sum_{1 \leqslant d \leqslant x} \sum_{1 \leqslant m \leqslant x/d} f(d).$$

Here, in the last summation, we made the change of variable $n = md$. Thus we obtain

$$\sum_{1 \leqslant n \leqslant x} g(n) = \sum_{1 \leqslant d \leqslant x} f(d) \sum_{1 \leqslant m \leqslant x/d} 1 = \sum_{1 \leqslant d \leqslant x} f(d) \left\lfloor \frac{x}{d} \right\rfloor,$$

where, as usual, we write $\lfloor \theta \rfloor$ for the greatest integer not exceeding θ. Thus we see that

$$\frac{1}{x} \sum_{1 \leqslant n \leqslant x} g(n) = \frac{1}{x} \sum_{1 \leqslant d \leqslant x} f(d) \left(\frac{x}{d} + O(1) \right) = \sum_{1 \leqslant d \leqslant x} \frac{f(d)}{d} + O\left(\frac{1}{x} \sum_{1 \leqslant d \leqslant x} |f(d)| \right).$$

In many circumstances, the first term on the right hand side of the last equation is of larger order of magnitude than the last term, and then one has the asymptotic formula

$$\frac{1}{x} \sum_{1 \leqslant n \leqslant x} g(n) \sim \sum_{1 \leqslant d \leqslant x} f(d)/d,$$

a formula that is useful provided that the new average is easier to compute than the original average. We illustrate these ideas with some examples.

Example 15.2. The Euler totient $\phi(n)$.

We obtained earlier the formula

$$\phi(n) = n \sum_{d|n} \mu(d)/d,$$

by using the Möbius inversion formulae. Following the above strategy, we find that

$$\begin{aligned}\sum_{1\leqslant n\leqslant x} \phi(n) &= \sum_{1\leqslant n\leqslant x} \sum_{d|n} \mu(d)n/d = \sum_{1\leqslant d\leqslant x} \mu(d) \sum_{1\leqslant m\leqslant x/d} m \\ &= \sum_{1\leqslant d\leqslant x} \mu(d) \cdot \tfrac{1}{2}\lfloor x/d\rfloor(\lfloor x/d\rfloor + 1) = \sum_{1\leqslant d\leqslant x} \mu(d) \left(\frac{x^2}{2d^2} + O(x/d)\right) \\ &= \tfrac{1}{2}x^2 \sum_{1\leqslant d\leqslant x} \mu(d)/d^2 + O\left(\sum_{1\leqslant d\leqslant x} |\mu(d)|x/d\right).\end{aligned}$$

But

$$\sum_{1\leqslant d\leqslant x} \mu(d)/d^2 = \sum_{d=1}^{\infty} \mu(d)/d^2 + O\left(\sum_{d>x} 1/d^2\right) = C + O(1/x),$$

where $C = \prod_p (1 - 1/p^2) = 1/\zeta(2) = 6/\pi^2$ (fact!). In addition, one has

$$\sum_{1\leqslant d\leqslant x} \frac{1}{d} < 1 + \int_1^x \frac{d\theta}{\theta} = O(\log x).$$

Thus

$$\frac{1}{x} \sum_{1\leqslant n\leqslant x} \phi(n) = \frac{3}{\pi^2}x + O(\log x).$$

In some sense, this means that the average order of $\phi(n)$ is $(6/\pi^2)n$ (why?). On the other hand, the upper bound $\phi(n) \leqslant n$ is trivial (and for every prime number p one has $\phi(p) = p - 1$). It is possible, though harder, to show that for all natural numbers n, the Euler totient $\phi(n)$ is asymptotically larger than $e^{-\gamma}n/\log\log n$, where $\gamma = 0.577\ldots$ is Euler's constant.

Example 15.3. The sum of divisors function $\sigma(n)$.

In this instance we have the formula $\sigma(n) = \sum_{d|n} d$, and so

$$\begin{aligned}\sum_{1\leqslant n\leqslant x} \sigma(n) &= \sum_{1\leqslant n\leqslant x} \sum_{d|n} n/d = \sum_{1\leqslant d\leqslant x} \sum_{1\leqslant m\leqslant x/d} m \\ &= \sum_{1\leqslant d\leqslant x} \tfrac{1}{2}\lfloor x/d\rfloor(\lfloor x/d\rfloor + 1) = \sum_{1\leqslant d\leqslant x} \left(\frac{x^2}{2d^2} + O(x/d)\right).\end{aligned}$$

But

$$\sum_{1\leqslant d\leqslant x} 1/d^2 = \sum_{d=1}^{\infty} 1/d^2 + O\left(\sum_{d>x} 1/d^2\right) = \zeta(2) + O(1/x),$$

and hence

$$\frac{1}{x} \sum_{1\leqslant n\leqslant x} \sigma(n) = \tfrac{1}{2}\zeta(2)x + O(\log x) = \frac{\pi^2}{12}x + O(\log x).$$

Example 15.4. The divisor function $\tau(n)$.

In this instance, of course, one has $\tau(n)=\sum_{d|n}1$, and so

$$\begin{aligned}\sum_{1\leqslant n\leqslant x}\tau(n)&=\sum_{1\leqslant n\leqslant x}\sum_{d|n}1=\sum_{1\leqslant d\leqslant x}\sum_{1\leqslant m\leqslant x/d}1\\&=\sum_{1\leqslant d\leqslant x}\lfloor x/d\rfloor=\sum_{1\leqslant d\leqslant x}(x/d+O(1))\\&=x\sum_{1\leqslant d\leqslant x}1/d+O(x).\end{aligned}$$

But (as a good exercise in calculus),

$$\begin{aligned}\sum_{1\leqslant d\leqslant x}1/d&=\sum_{1\leqslant d\leqslant x}\left(\int_{d-1/2}^{d+1/2}\frac{dt}{t}+O(1/d^2)\right)\\&=\log x+O(1).\end{aligned}$$

Thus we deduce that

$$\frac{1}{x}\sum_{1\leqslant n\leqslant x}\tau(n)=\log x+O(1).$$

Perhaps it is worth noting that one can refine the above formula to obtain

$$\sum_{1\leqslant n\leqslant x}\tau(n)=x\log x+(2\gamma-1)x+O(\sqrt{x}),$$

where $\gamma=0.577\dots$ is Euler's constant.

Example 15.5. The number of squarefree numbers.

We turn our attention next to counting the number of squarefree numbers up to x. Define

$$S(x)=\mathrm{card}\{1\leqslant n\leqslant x\,:\,n\text{ is squarefree}\}.$$

In order to analyse this sum, we need to have available a detector for squarefree numbers. If we recall that the sum of the Möbius function over the divisors of an integer n, which we called $\nu(n)$, is non-zero precisely when $n=1$, in which case it is 1, we are led to consider the expression

$$\sum_{d^2|n}\mu(d).$$

Let m be the largest positive integer with $m^2|n$. Then the above expression is

$$\sum_{d|m}\mu(d)=\nu(m)=\begin{cases}1, & \text{when } m \text{ is equal to 1, i.e. } n \text{ is squarefree},\\ 0, & \text{when } m>1, \text{ i.e. } n \text{ is not squarefree}.\end{cases}$$

Thus we find that

$$S(x)=\sum_{1\leqslant n\leqslant x}\sum_{d^2|n}\mu(d).$$

But this expression has similar shape to those that we have considered before in this section, and so we may analyse this sum similarly. By means of the change of variable $n = md^2$, one finds that

$$\begin{aligned} S(x) &= \sum_{1\leqslant d\leqslant\sqrt{x}} \sum_{1\leqslant m\leqslant x/d^2} \mu(d) = \sum_{1\leqslant d\leqslant\sqrt{x}} \mu(d)\lfloor x/d^2\rfloor \\ &= \sum_{1\leqslant d\leqslant\sqrt{x}} \mu(d)\left(\frac{x}{d^2}+O(1)\right) = x\sum_{1\leqslant d\leqslant\sqrt{x}} \frac{\mu(d)}{d^2} + O\left(\sum_{1\leqslant d\leqslant\sqrt{x}} 1\right). \end{aligned}$$

But

$$\sum_{1\leqslant d\leqslant\sqrt{x}} \mu(d)/d^2 = \sum_{d=1}^{\infty}\frac{\mu(d)}{d^2} + O\left(\sum_{d>\sqrt{x}} 1/d^2\right) = 1/\zeta(2) + O(1/\sqrt{x}).$$

Thus we conclude that

$$\begin{aligned} S(x) &= x\left(\frac{1}{\zeta(2)} + O(1/\sqrt{x})\right) + O(\sqrt{x}) \\ &= \frac{6}{\pi^2}x + O(\sqrt{x}). \end{aligned}$$

Thus the "probability" that a randomly chosen positive integer is squarefree is $6/\pi^2$. This provides the world's worst method of calculating π (or rather π^2). As an exercise, consider the problem of counting the number of cubefree integers up to x (those n satisfying the property that whenever m^3 divides n, then $m = 1$). What about k-free numbers? (those integers n satisfying the property that whenever $m^k|n$, then $m = 1$).

Example 15.6. The number of distinct prime divisors $\omega(n)$.

We observe that

$$\sum_{1\leqslant n\leqslant x} \omega(n) = \sum_{1\leqslant n\leqslant x}\sum_{p|n} 1 = \sum_{p\leqslant x}\sum_{\substack{1\leqslant n\leqslant x\\ p|n}} 1 = \sum_{p\leqslant x}\lfloor x/p\rfloor.$$

At the end of section 3 we sketched an argument (as an exercise) for proving that there are positive constants a and b with $0 < a < 1 < b$ for which one has

$$ax/\log x < \sum_{p\leqslant x} 1 < bx/\log x \quad (x \geqslant 2).$$

These inequalities show that there is a constant $c > 3$ having the property that, for all $y \geqslant 2$, the number of prime numbers between y and cy is at least $y/\log y$ and at most

$c^2y/\log(cy)$. Thus, we have

$$\begin{aligned}\sum_{1\leqslant n\leqslant x}\omega(n)&=x\sum_{p\leqslant x}1/p+O(x)\\&\leqslant x\sum_{1\leqslant j\leqslant 1+\log x}\ \sum_{c^{j-1}<p\leqslant c^j}\frac{1}{p}+O(x)\\&\leqslant x\sum_{1\leqslant j\leqslant 1+\log x}\frac{c^{j+1}/(j\log c)}{c^{j-1}}+O(x)\\&\leqslant c^2x\sum_{1\leqslant j\leqslant 1+\log x}1/j+O(x).\end{aligned}$$

Hence we deduce that when x is large enough, one has

$$\sum_{1\leqslant n\leqslant x}\omega(n)\leqslant c^2x\log\log x+O(x).$$

On the other hand,

$$\begin{aligned}\sum_{1\leqslant n\leqslant x}\omega(n)&=x\sum_{p\leqslant x}1/p+O(x)\\&\geqslant x\sum_{1\leqslant j\leqslant \log x}\ \sum_{c^{j-1}<p\leqslant c^j}\frac{1}{p}+O(x)\\&\geqslant x\sum_{2\leqslant j\leqslant \log x}\frac{c^{j-1}/((j-1)\log c)}{c^{j}}+O(x)\\&\geqslant c^{-2}x\sum_{2\leqslant j\leqslant \log x}1/(j-1)+O(x).\end{aligned}$$

Hence

$$\sum_{1\leqslant n\leqslant x}\omega(n)\geqslant c^{-2}x\log\log x+O(x).$$

Thus we deduce that

$$c^{-2}\log\log x+O(1)\leqslant\frac{1}{x}\sum_{1\leqslant n\leqslant x}\omega(n)\leqslant c\log\log x+O(1).$$

In fact, one knows that asymptotically, one has $a=b=1$, and so the average size of $\omega(n)$ between 1 and x is asymptotically $\log\log x$.

Example 15.7. An upper bound for $\tau(n)$.

We show that for any positive number ε, one has $\tau(n)=O(n^\varepsilon)$ for $n\in\mathbb{N}$. In order to establish this estimate, we exploit the multiplicative property of $\tau(n)$, and investigate the function

$$\frac{\tau(n)}{n^\varepsilon}=\prod_{p^j\|n}\frac{j+1}{p^{j\varepsilon}}.$$

If one has $\varepsilon j\log p>\log(j+1)$, then $(j+1)p^{-j\varepsilon}<1$, and moreover $(\log(j+1))/j$ is a decreasing function of j when $j>e$. Thus there exists a number $C=C(\varepsilon)$, depending at most on the choice of ε, satisfying the property that whenever $p > C(\varepsilon)$, one has $\varepsilon j \log p > \log(j+1)$. We therefore deduce that

$$\tau(n)/n^{\varepsilon} \leqslant \prod_{\substack{p^j \| n \\ p \leqslant C(\varepsilon)}} \frac{j+1}{p^{j\varepsilon}} \prod_{p > C(\varepsilon)} 1.$$

But since there are just finitely many primes not exceeding $C(\varepsilon)$ (in fact, fewer than $C(\varepsilon)$ of them), the values of $(j+1)p^{-j\varepsilon}$ are bounded above by some number $A(\varepsilon)$ for $p \leqslant C(\varepsilon)$. Consequently,

$$\tau(n)/n^{\varepsilon} \leqslant \prod_{\substack{p | n \\ p \leqslant C(\varepsilon)}} A(\varepsilon) \leqslant A(\varepsilon)^{C(\varepsilon)}.$$

We therefore see that there is a positive number $B = B(\varepsilon)$, depending at most on ε, for which $\tau(n) \leqslant B(\varepsilon) n^{\varepsilon}$ for every natural number n, that is, for each positive number ε, one has $\tau(n) = O(n^{\varepsilon})$.

16. Diophantine approximation

Many important ideas in Number Theory stem from notions of Diophantine approximation, which is to say rational approximations to real numbers with prescribed properties.

Theorem 16.1 (Dirichlet, 1842)**.** *Let $\theta \in \mathbb{R}$ and let Q be a real number exceeding 1. Then there exist integers p and q with $1 \leqslant q < Q$ and $(p,q) = 1$ such that $|q\theta - p| \leqslant 1/Q$.*

Proof. We apply the Box Principle. Write $N = \lceil Q \rceil$, and consider the $N+1$ real numbers

$$0,\ 1,\ \{\theta\},\ \{2\theta\},\ \ldots,\ \{(N-1)\theta\},$$

where here, and throughout, we write $\{x\}$ for $x - \lfloor x \rfloor$. These $N+1$ real numbers all lie in the interval $[0,1]$. But if we divide this unit interval into N disjoint intervals of length $1/N$, it follows that there must be two numbers from the above set which necessarily lie in the same interval. The difference between these two numbers has the shape $q\theta - p$, where p and q are integers with $0 < q < N$. Thus we deduce that there exist integers p and q with $1 \leqslant q < Q$ and $|q\theta - p| \leqslant 1/Q$. The coprimality condition is obtained easily by dividing through by (p,q). □

Corollary 16.2. *Whenever θ is irrational, there exist infinitely many distinct pairs $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $(p,q) = 1$ and $|\theta - p/q| < 1/q^2$.*

Proof. Let $Q \geqslant 2$. Then by Dirichlet's theorem on Diophantine approximation, there exist $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $(p,q) = 1$, $q < Q$ and $0 < |\theta - p/q| \leqslant 1/(qQ) < 1/q^2$. Let Q' be any real number exceeding $|\theta - p/q|^{-1}$. A second application of Dirichlet's theorem shows that there exist $p' \in \mathbb{Z}$ and $q' \in \mathbb{N}$ with $(p', q') = 1$, $1 \leqslant q' < Q'$ and

$$|\theta - p'/q'| \leqslant 1/(q'Q') < |\theta - p/q|/q' \leqslant |\theta - p/q|.$$

Further, one has $|\theta - p'/q'| < 1/(q')^2$. Thus, necessarily, one has $p'/q' \neq p/q$. By iterating this process, we obtain a sequence $(p_n/q_n)_{n=1}^{\infty}$ of rational numbers with

$$0 < |\theta - p_n/q_n| < |\theta - p_{n-1}/q_{n-1}| < \cdots < |\theta - p_1/q_1|,$$

and $|\theta - p_i/q_i| < 1/q_i^2$, and hence infinitely many approximations p/q with $(p,q) = 1$ and $|\theta - p/q| < 1/q^2$. □

The continued fraction algorithm This provides a bijective correspondence between:

$$\text{rational } \theta \leftrightarrow \text{(finite continued fractions) } (a_0, a_1, \dots, a_n),$$
$$n \text{ finite, } a_i \in \mathbb{N}\ (i \geqslant 1),\ a_0 \in \mathbb{Z},\ a_n \geqslant 2,$$

$$\text{irrational } \theta \leftrightarrow \text{(infinite continued fractions) } (a_0, a_1, \dots),$$
$$a_0 \in \mathbb{Z},\ a_i \in \mathbb{N}\ (i \geqslant 1).$$

Algorithm: Given $\theta \in \mathbb{R}$, define the integers a_j $(j \geqslant 0)$ as follows.
Let $a_0 = \lfloor \theta \rfloor$. If $a_0 = \theta$ then stop.
If $a_0 \neq \theta$, define θ_1 by means of $\theta = a_0 + 1/\theta_1$, so that $\theta_1 > 1$.
Let $a_1 = \lfloor \theta_1 \rfloor$. If $a_1 = \theta_1$ then stop.

$$\dots$$

At step n, we suppose that the integers $a_0, a_1, \dots, a_n$ have been defined, that $\theta_n \in \mathbb{R}$ has been defined, and that $a_n = \lfloor \theta_n \rfloor$. If $a_n = \theta_n$ then stop.
If $a_n \neq \theta_n$, define θ_{n+1} by means of $\theta_n = a_n + 1/\theta_{n+1}$, so that $\theta_{n+1} > 1$.
Let $a_{n+1} = \lfloor \theta_{n+1} \rfloor$. If $a_{n+1} = \theta_{n+1}$ then stop.

$$\dots$$

If this algorithm terminates, say with the sequence $(a_0, a_1, \dots, a_n)$, then θ is rational and

$$\theta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\dots + \cfrac{1}{a_n}}}}.$$

It will be apparent shortly that when θ is rational, then necessarily the Continued Fraction Algorithm terminates.

For the above expansion, it is usually more convenient to write

$$\theta = a_0 + \frac{1}{a_1+}\,\frac{1}{a_2+}\,\dots\,\frac{1}{a_{n-1}+}\,\frac{1}{a_n} \quad \text{or} \quad \theta = [a_0; a_1, \dots, a_n].$$

If the continued Fraction Algorithm does not terminate, so that $a_n \neq \theta_n$ for each natural number n, then it follows (as we will see) that θ is irrational, and θ may be written in the form

$$\theta = a_0 + \frac{1}{a_1+}\,\frac{1}{a_2+}\,\dots\,\frac{1}{a_{n-1}+}\,\dots \quad \text{or} \quad \theta = [a_0; a_1, a_2, \dots].$$

Example 16.3. Write 57/32 as a continued fraction.
Put $\theta = 57/32$. Then $a_0 = \lfloor \theta \rfloor = 1$, and

$$\theta_1 = \frac{1}{\frac{57}{32} - 1} = \frac{32}{25}.$$

Then $a_1 = \lfloor \theta_1 \rfloor = 1$, and

$$\theta_2 = \frac{1}{\frac{32}{25} - 1} = \frac{25}{7}.$$

Then $a_2 = \lfloor \theta_2 \rfloor = 3$, and

$$\theta_3 = \frac{1}{\frac{25}{7} - 3} = \frac{7}{4}.$$

Then $a_3 = \lfloor \theta_3 \rfloor = 1$, and

$$\theta_4 = \frac{1}{\frac{7}{4} - 1} = \frac{4}{3}.$$

Then $a_4 = \lfloor \theta_4 \rfloor = 1$, and

$$\theta_5 = \frac{1}{\frac{4}{3} - 1} = 3.$$

Then $a_5 = 3$ and $\theta_5 = a_5$, so stop.

In this way we find that $57/32 = [1; 1, 3, 1, 1, 3]$.

Example 16.4. Write $\sqrt{3}$ as a continued fraction.
Put $\theta = \sqrt{3}$. Then $a_0 = \lfloor \sqrt{3} \rfloor = 1$, and

$$\theta_1 = \frac{1}{\sqrt{3} - 1} = \tfrac{1}{2}(\sqrt{3} + 1).$$

Then $a_1 = \lfloor \theta_1 \rfloor = 1$, and

$$\theta_2 = \frac{1}{\frac{1}{2}(\sqrt{3} - 1)} = \sqrt{3} + 1.$$

Then $a_2 = \lfloor \theta_2 \rfloor = 2$, and

$$\theta_3 = \frac{1}{\sqrt{3} - 1} = \tfrac{1}{2}(\sqrt{3} + 1) = \theta_1,$$

and the sequence repeats.

In this way we find that $\sqrt{3} = [1; 1, 2, 1, 2, 1, 2, \dots]$, a periodic continued fraction that, by convention, we write as $[1; \overline{1, 2}]$.

Example 16.5. Find the continued fraction expansion of $\frac{1}{2}(10 - \sqrt{7})$.
Put $\theta = \frac{1}{2}(10 - \sqrt{7})$. Then $a_0 = \left\lfloor \frac{1}{2}(10 - \sqrt{7}) \right\rfloor = 3$, and

$$\theta_1 = \frac{1}{\frac{1}{2}(10 - \sqrt{7}) - 3} = \frac{2(4 + \sqrt{7})}{16 - 7} = \tfrac{1}{9}(8 + 2\sqrt{7}).$$

Then $a_1 = \lfloor \theta_1 \rfloor = 1$, and

$$\theta_2 = \frac{1}{\frac{1}{9}(8 + 2\sqrt{7}) - 1} = \frac{9(-1 - 2\sqrt{7})}{1 - 28} = \tfrac{1}{3}(1 + 2\sqrt{7}).$$

Then $a_2 = \lfloor \theta_2 \rfloor = 2$, and

$$\theta_3 = \frac{1}{\frac{1}{3}(1 + 2\sqrt{7}) - 2} = \frac{3(-5 - 2\sqrt{7})}{25 - 28} = 5 + 2\sqrt{7}.$$

Then $a_3 = \lfloor \theta_3 \rfloor = 10$, and

$$\theta_4 = \frac{1}{(5 + 2\sqrt{7}) - 10} = \frac{-5 - 2\sqrt{7}}{25 - 28} = \tfrac{1}{3}(5 + 2\sqrt{7}).$$

Then $a_4 = \lfloor \theta_4 \rfloor = 3$, and

$$\theta_5 = \frac{1}{\frac{1}{3}(5+2\sqrt{7})-3} = \frac{3(-4-2\sqrt{7})}{16-28} = \tfrac{1}{2}(2+\sqrt{7}).$$

Then $a_5 = \lfloor \theta_5 \rfloor = 2$, and

$$\theta_6 = \frac{1}{\frac{1}{2}(2+\sqrt{7})-2} = \frac{2(-2-\sqrt{7})}{4-7} = \tfrac{1}{3}(4+2\sqrt{7}).$$

Then $a_6 = \lfloor \theta_6 \rfloor = 3$, and

$$\theta_7 = \frac{1}{\frac{1}{3}(4+2\sqrt{7})-3} = \frac{3(-5-2\sqrt{7})}{25-28} = 5+2\sqrt{7} = \theta_3,$$

and the sequence repeats.

In this way we find that

$$\tfrac{1}{2}(10-\sqrt{7}) = [3;1,2,10,3,2,3,10,3,2,3,\dots] = [3;1,2,\overline{10,3,2,3}].$$

Definition 16.6. In the above description of the continued fraction algorithm, and the resulting continued fraction expansion of a real number θ, the integers a_i are known as the **partial quotients** of θ, the real numbers θ_i are known as the **complete quotients** of θ, and the rational numbers

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

where p_n and q_n are relatively prime integers with $q_n \geqslant 1$, are known as the **convergents** to θ.

Theorem 16.7. *Let a_n $(n \geqslant 0)$ be the partial quotients of a real number θ, let θ_n be the corresponding complete quotients, and p_n/q_n the associated convergents. Then the integers p_n and q_n satisfy the recurrence relations*

$$p_0 = a_0, \quad q_0 = 1, \quad p_1 = a_0a_1 + 1, \quad q_1 = a_1,$$

and for $n \geqslant 2$,

$$p_n = a_np_{n-1} + p_{n-2}, \quad q_n = a_nq_{n-1} + q_{n-2}.$$

Furthermore,

$$p_nq_{n+1} - p_{n+1}q_n = (-1)^{n+1},$$

and when θ has an infinite continued fraction expansion, one has $q_n \to \infty$ as $n \to \infty$, and $\lim_{n\to\infty} p_n/q_n = \theta$.

Proof. We begin by establishing the claimed recurrences. Observe first that

$$\frac{p_1}{q_1} = \frac{a_0a_1+1}{a_1} = a_0 + \frac{1}{a_1} = [a_0, a_1],$$

and

$$\frac{p_2}{q_2} = \frac{a_2p_1 + p_0}{a_2q_1 + q_0} = \frac{a_2(a_0a_1 + 1) + a_0}{a_2a_1 + 1} = a_0 + \frac{a_2}{a_2a_1 + 1}$$
$$= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2}} = [a_0; a_1, a_2].$$

Thus the desired recurrences hold for $n = 0, 1, 2$, in view of the relations

$$1 = (a_0, 1) = (a_0a_1 + 1, a_1) = (a_2, a_2a_1 + 1) = (a_0a_1a_2 + a_0 + a_2, a_2a_1 + 1),$$

that yield the coprimality conditions $1 = (p_0, q_0) = (p_1, q_1) = (p_2, q_2)$.

Suppose then that the claimed recurrences hold for $n \leqslant m - 1$, for a fixed m with $m \geqslant 3$. We consider the continued fraction expansion $[a_1; a_2, \dots, a_m]$, a rational number associated with $[a_0; a_1, \dots, a_m]$. For $j \geqslant 0$, define the integers p'_j and $q'_j \geqslant 1$ by means of the formula

$$\frac{p'_j}{q'_j} = [a_1; a_2, \dots, a_{j+1}] \quad \text{and} \quad (p'_j, q'_j) = 1.$$

Then by the inductive hypothesis, one has $p'_0 = a_1$, $q'_0 = 1$, $p'_1 = a_1a_2 + 1$, $q'_1 = a_2$,

$$p'_n = a_{n+1}p'_{n-1} + p'_{n-2} \quad \text{and} \quad q'_n = a_{n+1}q'_{n-1} + q'_{n-2} \quad (2 \leqslant n \leqslant m - 1).$$

But

$$[a_0; a_1, \dots, a_{j+1}] = a_0 + \frac{q'_j}{p'_j},$$

and hence

$$\frac{p_{j+1}}{q_{j+1}} = \frac{a_0p'_j + q'_j}{p'_j} \quad \text{for} \quad 0 \leqslant j \leqslant m - 1.$$

Thus, since $(a_0p'_j + q'_j, p'_j) = (q'_j, p'_j) = 1$ and $(p_{j+1}, q_{j+1}) = 1$, we obtain

$$p_{j+1} = a_0p'_j + q'_j \quad \text{and} \quad q_{j+1} = p'_j \quad \text{for} \quad 0 \leqslant j \leqslant m - 1.$$

But then the recurrence formulae for p'_j and q'_j imply that

$$\begin{aligned} p_m - &a_mp_{m-1} - p_{m-2} \\ &= a_0(p'_{m-1} - a_mp'_{m-2} - p'_{m-3}) + (q'_{m-1} - a_mq'_{m-2} - q'_{m-3}) \\ &= 0 \end{aligned}$$

and

$$q_m - a_mq_{m-1} - q_{m-2} = p'_{m-1} - a_mp'_{m-2} - p'_{m-3} = 0.$$

Then the inductive hypothesis holds with $m + 1$ in place of m, and thus the desired recurrence relations hold.

We prove the relation $p_nq_{n+1} - p_{n+1}q_n = (-1)^{n+1}$ by induction, noting that

$$p_0q_1 - p_1q_0 = a_0a_1 - (a_0a_1 + 1) = -1$$

and

$$\begin{aligned} p_nq_{n+1} - p_{n+1}q_n &= p_n(a_{n+1}q_n + q_{n-1}) - q_n(a_{n+1}p_n + p_{n-1}) \\ &= -(p_{n-1}q_n - p_nq_{n-1}). \end{aligned}$$

Suppose now that θ has an infinite continued fraction expansion. Then, in view of the positivity of the integers a_i for $i \geqslant 1$, it follows from the relation $q_n = a_n q_{n-1} + q_{n-2}$ that $q_n \geqslant q_{n-1} + q_{n-2}$ for $n \geqslant 2$, whence $q_n \to \infty$ as $n \to \infty$.

Finally, since $p_n q_{n+1} - p_{n+1} q_n = (-1)^{n+1}$, we find that

$$\left| \frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}} \right| = \frac{1}{q_n q_{n+1}}.$$

But $\theta = [a_0; a_1, \dots, a_n, \theta_{n+1}]$, where $0 < 1/\theta_{n+1} \leqslant 1/a_{n+1}$. Thus we find that θ lies between p_n/q_n and p_{n+1}/q_{n+1} for each natural number n, whence

$$|\theta - p_n/q_n| \leqslant 1/(q_n q_{n+1}).$$

In particular, it follows that $\lim_{n\to\infty} p_n/q_n = \theta$. □

Suppose that $\theta = s/t$ with $(s,t) = 1$ and $t \geqslant 1$, which is to say that θ is rational. Then for every convergent p_n/q_n, one has either $p_n/q_n = s/t$, or else

$$\frac{1}{tq_n} \leqslant \left| \frac{s}{t} - \frac{p_n}{q_n} \right| \leqslant \frac{1}{q_n q_{n+1}}.$$

Since $q_{n+1} > t$ for n sufficiently large, it follows that there is a natural number n with $\theta = p_n/q_n$. Thus rational numbers possess terminating continued fraction expansions.

It is a fact that θ is a quadratic irrational number if and only if its continued fraction expansion is ultimately periodic. Meanwhile, certain real numbers have continued fraction expansions that are elegant to describe, such as

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, \dots].$$

In the above examples we have noted that the continued fraction expansions of quadratic irrational numbers seem to have special properties (periodic continued fraction expansions). Motivated by this observation, we now discuss rational approximations to algebraic numbers.

Definition 16.8. We say that the real number θ is **algebraic** and **of degree** d if there exists a polynomial $f(t) \in \mathbb{Z}[t]$ such that (i) $\deg(f) = d$, (ii) f is irreducible over $\mathbb{Q}$, and (iii) one has $f(\theta) = 0$.

Note that the degree d of θ is unique, for if f and g are distinct polynomials for which $f(\theta) = g(\theta) = 0$, then by the division algorithm for polynomials, there is some greatest common divisor h of f and g for which $h(\theta) = 0$. But then f and g cannot both be irreducible.

Definition 16.9. We say that the real number θ is **transcendental** if θ is **not** algebraic of any degree.

Recall that an argument based on countability shows that not all real numbers are algebraic, and indeed that almost all real numbers are transcendental. However, it was not until 1844 that any explicit transcendental number was exhibited — or indeed that transcendental numbers were known to exist at all.

Given an algebraic number α of degree d, consider the set of all polynomials $f(t) \in \mathbb{Z}[t]$ for which $f(\alpha) = 0$. As already noted above, by using the division algorithm for polynomials, it follows that there exists a unique minimal degree for non-trivial members of this set. If we then restrict to polynomials wherein the set of coefficients have no common factor, and the leading coefficient is positive, then we obtain a unique polynomial $p_\alpha(t) \in \mathbb{Z}[t]$, known as the **minimal** polynomial of α.

Theorem 16.10 (Liouville, 1844). *Suppose that α is an algebraic number of degree $n > 1$. Then there exists a positive constant $c = c(\alpha)$ such that whenever $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, one has $|\alpha - p/q| \geqslant c/q^n$.*

Proof. We may plainly suppose that $\alpha \in \mathbb{R}$, for otherwise the conclusion of the theorem is immediate. Write $f(t)$ for the minimal polynomial of α, so that $f(t) \in \mathbb{Z}[t]$ has degree n. Then by the Mean Value Theorem, given $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, there exists a real number ξ with ξ lying between α and p/q, such that

$$f(\alpha) - f(p/q) = (\alpha - p/q)f'(\xi).$$

But by hypothesis, the number α is irrational, and so $f(p/q) \neq 0$. We therefore see that

$$|q^n f(p/q)| \geqslant 1.$$

Moreover, since without loss of generality we may suppose that $|\alpha - p/q| < 1$, we find that $|\xi| < |\alpha| + 1$, and hence

$$|f'(\xi)| \leqslant \sup_{|z|<|\alpha|+1} |f'(z)|.$$

Writing $c(\alpha)^{-1}$ for the latter supremum, we conclude that

$$1/q^n \leqslant |f(p/q)| = |f(\alpha) - f(p/q)| = |\alpha - p/q| \cdot |f'(\xi)| \leqslant c(\alpha)^{-1}|\alpha - p/q|,$$

whence

$$|\alpha - p/q| \geqslant c(\alpha)/q^n.$$

□

It is worthwhile noting a simple enhancement of Liouville's theorem that is of utility in applications.

Theorem 16.11. *Suppose that α is a non-zero algebraic number of degree $n \geqslant 1$. Then there exists a positive constant $c = c(\alpha)$ such that whenever $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(p, q) = 1$, and q is sufficiently large, one has $|\alpha - p/q| \geqslant c/q^n$.*

Proof. When α is algebraic of degree exceeding 1, the desired conclusion is immediate from Liouville's theorem. It remains only to consider the situation in which α is rational, say $\alpha = r/s$ for some $r \in \mathbb{Z}$ and $s \in \mathbb{N}$ with $(r, s) = 1$. But then, whenever $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(p, q) = 1$, and q is larger than s, one has $r/s \neq p/q$, and so

$$|\alpha - p/q| = \left|\frac{r}{s} - \frac{p}{q}\right| = \left|\frac{qr - ps}{qs}\right| \geqslant \frac{1}{qs}.$$

Thus, when the degree of α is 1, the desired conclusion holds with $c(\alpha) = 1/(2s)$. □

Corollary 16.12. *Let $\theta = \sum_{n=1}^{\infty} 2^{-n!}$. Then θ is transcendental.*

Proof. Suppose that θ is algebraic of some degree $d \geqslant 1$. Then by Theorem 16.11 there exists a constant $c = c(\theta) > 0$ such that whenever $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(p, q) = 1$, one has

$$|\theta - p/q| > c(\theta)/q^d.$$

For each natural number j, write

$$p_j = 2^{j!} \sum_{n=1}^{j} 2^{-n!} \quad \text{and} \quad q_j = 2^{j!}.$$

Since p_j is odd, we have $(p_j, q_j) = 1$. Thus

$$|\theta - p_j/q_j| = \sum_{n=j+1}^{\infty} 2^{-n!} < 2^{1-(j+1)!} \leqslant 2^{-j \cdot j!} = q_j^{-j}.$$

Thus there exist infinitely many $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $(p, q) = 1$ and satisfying the property that $|\theta - p/q| < c(\theta)q^{-(d+1)}$ (just take j large enough that $j > d + 1$ and $q_j = 2^{j!} > c(\theta)^{-1}$), contradicting our earlier consequence of Theorem 16.11. This contradiction confirms that θ cannot be algebraic, and consequently is transcendental. □

In fact one can show that whenever $a \geqslant 2$ and $b \geqslant 3$ are integers, then the number $\sum_{n=1}^{\infty} a^{-b^n}$ is transcendental. It is also known that π is transcendental (Lindemann, 1882), and that e is transcendental (Hermite, 1873). Indeed, Lindemann proved that whenever $\alpha_1, \dots, \alpha_n$ are distinct algebraic numbers, and $\beta_1, \dots, \beta_n$ are non-zero algebraic numbers, then $\beta_1 e^{\alpha_1} + \cdots + \beta_n e^{\alpha_n} \neq 0$. Since $e^{i\pi} + 1 = 0$, it follows that π cannot be algebraic.

We note also the theorem of Gelfond-Schneider (1934) that resolved Hilbert's 7th problem: whenever $\alpha \neq 0, 1$ is algebraic, and β is algebraic and irrational, the number α^β is transcendental. Thus, for example, one sees that $2^{\sqrt{2}}$ and $e^\pi = (-1)^{-i}$ are both transcendental.

Open Problem: Is it true that e and π are algebraically independent? That is to say, is it true that there is **no** non-trivial polynomial $F(x, y) \in \mathbb{Z}[x, y]$ with the property that $F(e, \pi) = 0$?

17. Diophantine equations: Pell's equation

We investigate the solubility of the equation

$$x^2 - dy^2 = 1,$$

for a fixed integer d that is not a perfect square, in integers x and y. This turns out to be a topic intimately connected with the continued fraction expansion of $\sqrt{d}$. We note that the equation $x^2 - dy^2 = 1$ always has the (trivial) solutions $(x, y) = \pm(1, 0)$, so the relevant problem is of determining whether there are additional solutions, and how many such solutions exist.

Theorem 17.1. *Suppose that $d > 0$ is not a perfect square. Then the Diophantine equation $x^2 - dy^2 = 1$ has infinitely many solutions.*

Proof. We consider Diophantine approximations to $\sqrt{d}$. By Dirichlet's theorem on Diophantine approximation, for each integer $Q > 1$ there exist $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $1 \leqslant q < Q$ such that $|p - q\sqrt{d}| \leqslant 1/Q$. Given any such approximation, we have

$$|p + q\sqrt{d}| \leqslant Q^{-1} + 2q\sqrt{d} \leqslant 3q\sqrt{d} < 3Q\sqrt{d}.$$

Thus we deduce that

$$|p^2 - dq^2| = |(p - q\sqrt{d})(p + q\sqrt{d})| < 3\sqrt{d}.$$

Since $\sqrt{d}$ is irrational, moreover, there are infinitely many pairs (p, q) with this property, and hence infinitely many pairs (p, q) for which $p^2 - dq^2$ takes the same fixed value (for there are at most $6\sqrt{d}+1$ available values). Suppose then that $p^2 - dq^2 = N$ has infinitely many integral solutions. Note that since $\sqrt{d}$ is irrational, then $N \neq 0$. Since there are infinitely many solutions to the aforementioned equation, we may select two solutions, say (p, q) and (p', q'), with the property that $p \neq \pm p'$ and $q \neq \pm q'$, and $p \equiv p' \pmod{|N|}$ and $q \equiv q' \pmod{|N|}$. But then we have

$$\begin{aligned}(pp' - dqq')^2 - d(pq' - p'q)^2 &= (pp')^2 + d^2(qq')^2 - d(pq')^2 - d(p'q)^2 \\ &= (p^2 - dq^2)((p')^2 - d(q')^2) = N^2,\end{aligned}$$

whence

$$x = (pp' - dqq')/N, \quad y = (pq' - p'q)/N$$

provides a non-trivial integral solution of the equation $x^2 - dy^2 = 1$. In order to check that $(x, y) \neq (\pm 1, 0)$, observe that if $x = \pm 1$ and $y = 0$, then

$$pp' = dqq' \pm N \quad \text{and} \quad pq' = p'q,$$

whence

$$p^2p' = dq(pq') \pm pN = p'dq^2 \pm pN,$$

and so

$$p'(p^2 - dq^2) = \pm pN \quad \Rightarrow \quad Np' = \pm Np \quad \Rightarrow \quad p = \pm p' \quad \text{and} \quad q = \pm q'.$$

The latter contradicts our earlier hypothesis, and hence shows that $(x, y) \neq (\pm 1, 0)$.

Given this single non-trivial solution (x, y) of $x^2 - dy^2 = 1$, we generate infinitely many others by noting that whenever (u, v) is any one solution, then

$$(u^2 + dv^2)^2 - d(2uv)^2 = (u^2 - dv^2)^2 = 1,$$

whence $(u^2 + dv^2, 2uv)$ is a second solution with larger x-coordinate. By iterating this process we plainly obtain infinitely many distinct non-trivial solutions. □

Definition 17.2. There is a unique solution (x, y) of the equation $x^2 - dy^2 = 1$ in which x and y have their smallest positive values. This solution is called the **fundamental solution**.

Theorem 17.3. *Suppose that $d > 0$ is not a perfect square. Let (x_1, y_1) be the fundamental solution to the equation $x^2 - dy^2 = 1$. Then the equation $x^2 - dy^2 = 1$ has its solutions given by*

$$x + \sqrt{d}y = \pm(x_1 + \sqrt{d}y_1)^n \quad (n \in \mathbb{Z}).$$

Proof. We begin by associating to each solution (x,y) of $x^2-dy^2=1$ the real number $x+y\sqrt{d}$. Observe that given two such solutions, say (x,y) and (u,v), and the associated real numbers $\varepsilon=x+y\sqrt{d}$ and $\varepsilon'=u+v\sqrt{d}$, the numbers $\varepsilon\varepsilon'$ and ε/ε' also yield solutions. For we have

$$\varepsilon\varepsilon'=(x+y\sqrt{d})(u+v\sqrt{d})=(xu+dvy)+\sqrt{d}(uy+vx),$$

and

$$(xu+dvy)^2-d(uy+vx)^2=(x^2-dy^2)(u^2-dv^2)=1,$$

and similarly,

$$\varepsilon/\varepsilon'=\frac{(x+y\sqrt{d})(u-v\sqrt{d})}{u^2-dv^2}=(xu-dvy)+\sqrt{d}(-xv+uy),$$

and

$$(xu-dvy)^2-d(uy-vx)^2=(x^2-dy^2)(u^2-dv^2)=1.$$

Write now $\varepsilon=x_1+y_1\sqrt{d}$ for the real number corresponding to the fundamental solution (x_1,y_1). Let (x',y') be any other solution of the equation $x^2-dy^2=1$ with $(x',y')\neq(\pm1,0)$, and write $\varepsilon'=x'+y'\sqrt{d}$. If $0<\varepsilon'<1$, then we may consider instead the solution associated with the real number $x'-y'\sqrt{d}=1/(x'+y'\sqrt{d})$. In addition, if $\varepsilon'<0$, we instead consider the solution associated with the real number $-x'-y'\sqrt{d}$. In this way, we find that there is no loss of generality in supposing that $\varepsilon'>1$.

Next observe that since $\varepsilon>1$, then for some natural number m we have $\varepsilon^m\leqslant\varepsilon'<\varepsilon^{m+1}$. In view of our initial discussion, therefore, there is a solution (r,s) corresponding to the real number $\varepsilon'/\varepsilon^m$. But $1\leqslant\varepsilon'/\varepsilon^m<\varepsilon$, so from our hypothesis that ε corresponds to the fundamental solution, and so is minimal amongst all positive solutions, we deduce that $\varepsilon'=\varepsilon^m$. Thus we find that by our earlier discussion, one has $x'+y'\sqrt{d}=\pm(x_1+\sqrt{d}y_1)^m$ for some integer m. □

Continued fractions and Pell's equation The solutions of Pell's equation have a natural interpretation in terms of the convergents to the continued fraction expansion of $\sqrt{d}$.

Example 17.4. Recall that $\sqrt{3}=[1;\overline{1,2}]$. We consider convergents p_n/q_n to $\sqrt{3}$, and the corresponding values of $p_n^2-3q_n^2$. It is useful in this context to recall that when $\theta=[a_0;a_1,\dots]$, then the convergents p_n/q_n to θ satisfy the relations

$$p_0=a_0,\quad q_0=1,\quad p_1=a_0a_1+1,\quad q_1=a_1,$$

$$p_n=a_np_{n-1}+p_{n-2},\quad q_n=a_nq_{n-1}+q_{n-2}.$$

In the case at hand, we obtain

$$p_0=1 \text{ and } q_0=1\Rightarrow p_0^2-3q_0^2=-2,$$
$$p_1=1\cdot1+1=2 \text{ and } q_1=1\Rightarrow p_1^2-3q_1^2=4-3=1,$$
$$p_2=2p_1+p_0=5 \text{ and } q_2=2q_1+q_0=3\Rightarrow p_2^2-3q_2^2=25-3\cdot9=-2,$$
$$p_3=p_2+p_1=7 \text{ and } q_3=q_2+q_1=4\Rightarrow p_3^2-3q_3^2=49-3\cdot16=1,$$

and so on. So we find that for each natural number n, the pair $(x,y)=(p_{2n-1},q_{2n-1})$ provides a solution of the equation $x^2-3y^2=1$. Thus the general solution of the equation $x^2-3y^2=1$ is provided by the relation $x+y\sqrt{3}=\pm(2+\sqrt{3})^m$ $(m\in\mathbb{Z})$.

18. Pell's equation: the structure of solutions in detail.

We analyse the connection between continued fraction expansions of $\sqrt{d}$, and solutions of the Pell equation $x^2-dy^2=1$, in some detail.

Lemma 18.1. *The continued fraction $[a_0;a_1,a_2,\dots]$ represents a quadratic irrational number if and only if the sequence $(a_0,a_1,\dots)$ is ultimately periodic.*

Proof. We first show that when $\theta=[a_0,a_1,\dots,a_{k-1},\overline{a_k,a_{k+1},\dots,a_{k+m-1}}]$, then θ is a quadratic irrational number. Write $\phi=[\overline{a_k;a_{k+1},\dots,a_{k+m-1}}]$. Then we have

$$\phi=[a_k;a_{k+1},\dots,a_{k+m-1},\phi].$$

If we write p'_M/q'_M for the Mth convergent to ϕ, then

$$p'_0=a_k,\quad q'_0=1,\quad p'_1=a_ka_{k+1}+1,\quad q'_1=a_{k+1},$$

and we have

$$p'_M=a_{k+M}p'_{M-1}+p'_{M-2},\quad q'_M=a_{k+M}q'_{M-1}+q'_{M-2}\quad(2\leqslant M\leqslant m-1),$$

with

$$\frac{p'_M}{q'_M}=\frac{p'_{M-1}a_{k+M}+p'_{M-2}}{q'_{M-1}a_{k+M}+q'_{M-2}}=[a_k;a_{k+1},\dots,a_{k+M}]\quad(2\leqslant M\leqslant m-1).$$

The recurrences also apply for non-integral coefficients, so that

$$\phi=[a_k;a_{k+1},\dots,a_{k+m-1},\phi]=\frac{p'_{m-1}\phi+p'_{m-2}}{q'_{m-1}\phi+q'_{m-2}}.$$

Thus we obtain $q'_{m-1}\phi^2+(q'_{m-2}-p'_{m-1})\phi-p'_{m-2}=0$, whence ϕ is quadratic irrational. We next write p_M/q_M for the Mth convergent to θ, and deduce similarly that

$$\theta=[a_0,a_1,\dots,a_{k-1},\phi]=\frac{p_{k-1}\phi+p_{k-2}}{q_{k-1}\phi+q_{k-2}}.$$

But ϕ is quadratic irrational, and so θ must also be quadratic irrational, as can be confirmed from the formula

$$\frac{\alpha+\beta\sqrt{d}}{\gamma+\delta\sqrt{d}}=\frac{(\alpha+\beta\sqrt{d})(\gamma-\delta\sqrt{d})}{\gamma^2-d\delta^2}=\frac{(\alpha\gamma-d\beta\delta)+(\beta\gamma-\alpha\delta)\sqrt{d}}{\gamma^2-d\delta^2}.$$

Conversely, suppose that θ is a quadratic irrational, so that it satisfies an equation of the shape $a\theta^2+b\theta+c=0$ with $a,b,c\in\mathbb{Z}$ subject to the condition that $d=b^2-4ac>0$ is not a square. Write $f(x,y)=ax^2+bxy+cy^2$, and note that one then has $f(x,y)=(x,y)A(x,y)^T$, where

$$A=\begin{pmatrix} a & \frac{1}{2}b\\ \frac{1}{2}b & c\end{pmatrix}$$

is a matrix with $\det(A) = ac - \frac{1}{4}b^2 = -\frac{1}{4}d$. The *discriminant* of f is $d = -4\det(A)$. Suppose that p_n/q_n is the nth convergent to θ. Let

$$\gamma = \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix},$$

so that $\det(\gamma) = p_n q_{n-1} - p_{n-1}q_n = (-1)^{n+1}$. Then γ takes f to a quadratic form

$$f_n(x,y) = f(p_n x + p_{n-1}y, q_n x + q_{n-1}y) = a_n x^2 + b_n xy + c_n y^2,$$

say, having the same discriminant as f, since

$$\det(\gamma^T A\gamma) = (\det(\gamma))^2 \det(A) = \det(A).$$

Moreover, one has

$$a_n = f(p_n, q_n) \quad \text{and} \quad c_n = f(p_{n-1}, q_{n-1}) = a_{n-1}.$$

But $f(\theta, 1) = 0$, and we have $f(p_n/q_n, 1) = f(p_n, q_n)/q_n^2 = a_n/q_n^2$, whence

$$\begin{aligned} f(p_n/q_n, 1) &= f(p_n/q_n, 1) - f(\theta, 1) \\ &= a((p_n/q_n)^2 - \theta^2) + b(p_n/q_n - \theta). \end{aligned}$$

Thus we see that

$$|a_n| = q_n^2 |\theta - p_n/q_n| \cdot |a(\theta + p_n/q_n) + b|.$$

But $|\theta - p_n/q_n| \leqslant 1/(q_n q_{n+1}) < 1/q_n^2$, and thus

$$|a_n| < |a|(2|\theta| + |p_n/q_n - \theta|) + |b| \leqslant |a|(2|\theta| + 1) + |b|.$$

We therefore deduce that $|a_n|$ is bounded independently of n. Since $c_n = a_{n-1}$, the same conclusion holds also for c_n. Moreover, one has $b_n^2 - 4a_n c_n = d$, and so $|b_n| \leqslant \sqrt{d + 4a_n c_n}$ is also bounded independently of n.

Next write θ_n for the complete quotients of θ. Then for $n \geqslant 1$, one has

$$\theta = \frac{p_n\theta_{n+1} + p_{n-1}}{q_n\theta_{n+1} + q_{n-1}},$$

whence

$$\begin{aligned} f_n(\theta_{n+1}, 1) &= f(p_n\theta_{n+1} + p_{n-1}, q_n\theta_{n+1} + q_{n-1}) \\ &= (q_n\theta_{n+1} + q_{n-1})^2 f(\theta, 1) = 0. \end{aligned}$$

Since there are only finitely many available choices for the coefficients of the polynomial f_n, it follows that there are only finitely many possible choices for θ_n. Hence, for some positive numbers l and m, one has $\theta_{l+m} = \theta_l$, and so the continued fraction expansion of θ is ultimately periodic. □

Fact 18.2. $\sqrt{d}$ has continued fraction expansion of the shape

$$[a_0; \overline{a_1, a_2, a_3, \ldots, a_3, a_2, a_1, 2a_0}].$$

Lemma 18.3. *When d is a positive integer that is not a square, the continued fraction expansion of $\sqrt{d}$ takes the form $[a_0; \overline{a_1, \ldots, a_m}]$.*

Proof. The conclusion of the lemma follows from the assertion that $1/(\sqrt{d}-\lfloor\sqrt{d}\rfloor)$ has purely periodic continued fraction expansion. The latter is a consequence of the general conclusion that θ has purely periodic continued fraction expansion, when θ is quadratic irrational, provided that θ and the conjugate θ' of θ satisfy $\theta > 1$ and $-1 < \theta' < 0$. For suppose that $\theta = [a_0; a_1, \dots]$. Then since $\theta_n = a_n + 1/\theta_{n+1}$, by conjugation we obtain the relation $\theta'_n = a_n + 1/\theta'_{n+1}$. An inductive argument now shows that $-1 < \theta'_n < 0$ for $n \geqslant 0$. In order to confirm this assertion, observe first that $\theta'_0 = \theta'$ satisfies the claimed inequalities. Next, since $a_n \geqslant 1$ for all n, should one have $-1 < \theta'_n < 0$, then

$$-1 < \frac{1}{-a_n + \theta'_n} < 0,$$

and thus $-1 < \theta'_{n+1} < 0$. The desired conclusion does indeed therefore follow by induction. But θ is a quadratic irrational, so has ultimately periodic continued fraction expansion as a consequence of Lemma 18.1. Then one has $\theta_i = \theta_j$ for some $j > i$, and so likewise $1/\theta'_i = 1/\theta'_j$. But then one has

$$\begin{aligned}|a_{i-1} - a_{j-1}| &= |(\theta'_{i-1} - \theta'_{j-1}) - (1/\theta'_i - 1/\theta'_j)| \\ &= |\theta'_{i-1} - \theta'_{j-1}| < 1,\end{aligned}$$

so that $a_{i-1} = a_{j-1}$. It follows that $\theta_{i-1} = \theta_{j-1}$, whence by repeating this argument one deduces that $\theta = \theta_0 = \theta_{j-i}$. In this way we find that θ has purely periodic continued fraction expansion. □

Theorem 18.4. *Suppose that d is a positive integer which is not a perfect square, and that the continued fraction expansion $[a_0; a_1, \dots]$ of $\sqrt{d}$ has convergents p_n/q_n, and is ultimately periodic with period m. Then the only positive solutions (x, y) to the Pell equation $x^2 - dy^2 = 1$ are given by $(x, y) = (p_n, q_n)$, where $n = lm - 1$ for some natural number l, with n restricted to odd values.*

Proof. We begin by noting that whenever (x, y) is a solution to the equation $x^2 - dy^2 = 1$ with $x > 0$ and $y > 0$, then necessarily x/y is a convergent to the continued fraction expansion of $\sqrt{d}$. For we have

$$x^2 - dy^2 = 1 \quad \Rightarrow \quad x - y\sqrt{d} = 1/(x + y\sqrt{d}) > 0,$$

whence $x > y\sqrt{d}$ and $0 < x - y\sqrt{d} < 1/(2y\sqrt{d})$. Thus we deduce that

$$\left|\sqrt{d} - x/y\right| < 1/(2y^2).$$

We claim that whenever $\theta \in \mathbb{R}$ and x/y is a rational number with $(x, y) = 1$ satisfying $|\theta - x/y| < 1/(2y^2)$, then x/y is necessarily a convergent to the continued fraction expansion of θ. In order to establish this claim, recall that whenever p_n/q_n and p_{n+1}/q_{n+1} are successive convergents, then

$$p_n q_{n+1} - p_{n+1} q_n = (-1)^{n+1} \quad \Rightarrow \det\begin{pmatrix} p_n & p_{n+1} \\ q_n & q_{n+1}\end{pmatrix} = (-1)^{n+1}.$$

Then the equations

$$\begin{aligned} up_n + vp_{n+1} &= x \\ uq_n + vq_{n+1} &= y \end{aligned}$$

possess an integral solution (u, v). Note that if $u = 0$, then the coprimality of x and y ensures that $v = \pm 1$, whence $x/y = p_{n+1}/q_{n+1}$ is a convergent to θ, and similarly if $v = 0$. Thus we may suppose that $u \neq 0$ and $v \neq 0$. Since $q_n \to \infty$ as $n \to \infty$, moreover, we may choose n so that $q_n \leqslant y < q_{n+1}$. But $uq_n + vq_{n+1} = y$, and hence u and v must have opposite signs. Thus, using the fact that $q_n\theta - p_n$ and $q_{n+1}\theta - p_{n+1}$ have opposite signs, we deduce that

$$|y\theta - x| = |u(q_n\theta - p_n) + v(q_{n+1}\theta - p_{n+1})| \geqslant |q_n\theta - p_n|.$$

But since $|\theta - x/y| < 1/(2y^2)$, it follows that

$$\begin{aligned}\left|\frac{x}{y} - \frac{p_n}{q_n}\right| &\leqslant \left|\theta - \frac{x}{y}\right| + \left|\theta - \frac{p_n}{q_n}\right| \\ &\leqslant \left(\frac{1}{y} + \frac{1}{q_n}\right)|y\theta - x| \\ &< \frac{2}{q_n} \cdot \frac{1}{2y} = \frac{1}{q_n y},\end{aligned}$$

wherein we made use of the inequality $y \geqslant q_n$. But the latter inequality ensures that $x/y = p_n/q_n$ is a convergent to θ, completing the proof of our claim.

Now let p_n/q_n be the convergents to the continued fraction expansion of $\sqrt{d}$, and let θ_n be the corresponding complete quotients. Then

$$\sqrt{d} = \frac{p_n\theta_{n+1} + p_{n-1}}{q_n\theta_{n+1} + q_{n-1}}, \tag{18.1}$$

whence

$$\begin{aligned}q_n\sqrt{d} - p_n &= \frac{-p_n(q_n\theta_{n+1} + q_{n-1}) + q_n(p_n\theta_{n+1} + p_{n-1})}{q_n\theta_{n+1} + q_{n-1}} \\ &= \frac{p_{n-1}q_n - p_nq_{n-1}}{q_n\theta_{n+1} + q_{n-1}} = \frac{(-1)^n}{q_n\theta_{n+1} + q_{n-1}}.\end{aligned}$$

It follows that whenever n is even, one has $q_n\sqrt{d} > p_n$, whence

$$p_n^2 - dq_n^2 = (p_n - q_n\sqrt{d})(p_n + q_n\sqrt{d}) < 0.$$

Then $p_n^2 - dq_n^2 = 1$ can be satisfied only when n is odd.

Next, on recalling (18.1), we find that

$$(p_n - q_n\sqrt{d})\theta_{n+1} = q_{n-1}\sqrt{d} - p_{n-1},$$

whence

$$\begin{aligned}(p_n^2 - dq_n^2)\theta_{n+1} &= (p_n + q_n\sqrt{d})(q_{n-1}\sqrt{d} - p_{n-1}) \\ &= (p_nq_{n-1} - p_{n-1}q_n)\sqrt{d} + (dq_nq_{n-1} - p_np_{n-1}) \\ &= (-1)^{n+1}\sqrt{d} + N,\end{aligned}$$

for a suitable integer N. It follows that if $p_n^2 - dq_n^2 = 1$, so that necessarily n is odd, then one has $\theta_{n+1} = \sqrt{d} + N$, for some integer N. We now make use of the fact that the continued fraction expansion of $\sqrt{d}$ takes the form $[a_0; \overline{a_1, \ldots, a_m}]$, a conclusion that we established in Lemmata 18.1 and 18.3 above. One therefore has

$$\sqrt{d} = a_0 + 1/\theta_1,$$

where θ_1 has purely periodic continued fraction expansion. Thus we have

$$\theta_{n+1} = \sqrt{d} + N \quad \text{and} \quad \theta_{n+1} = a_{n+1} + 1/\theta_{n+2},$$

and $\sqrt{d} = a_0 + 1/\theta_1$, where $\theta_1 > 1$ and $\theta_{n+2} > 1$. Consequently,

$$a_{n+1} + 1/\theta_{n+2} = \sqrt{d} + N = a_0 + N + 1/\theta_1 \quad \Rightarrow \quad a_{n+1} = a_0 + N,$$

and thus $\theta_{n+2} = \theta_1$. But θ_1 is purely periodic with period m, and thus $m|(n+1)$. So if $p_n^2 - dq_n^2 = 1$, then n is odd and $n = lm - 1$ for some natural number l.

Thus far we have given necessary conditions for (p_n, q_n) to give a solution. We now show that these conditions are sufficient. Suppose then that $n = lm - 1$ is odd. Then by the periodicity of the continued fraction expansion of θ, one has $\theta_1 = \theta_{n+2}$. Consequently,

$$\sqrt{d} = \frac{p_{n+1}\theta_1 + p_n}{q_{n+1}\theta_1 + q_n}.$$

But $\sqrt{d} = a_0 + 1/\theta_1$, whence $\theta_1 = 1/(\sqrt{d} - a_0)$, and on substitution we obtain

$$\sqrt{d}(q_{n+1} + q_n(\sqrt{d} - a_0)) = p_{n+1} + p_n(\sqrt{d} - a_0),$$

whence on equating coefficients of $\sqrt{d}$ and 1, we obtain the relations

$$q_{n+1} - a_0 q_n = p_n \quad \text{and} \quad dq_n = p_{n+1} - a_0 p_n.$$

On eliminating a_0, we deduce that

$$p_n^2 - dq_n^2 = p_n q_{n+1} - p_{n+1} q_n = (-1)^{n+1} = 1,$$

on noting that n is presumed to be odd. Thus $(x, y) = (p_n, q_n)$ is indeed a solution of $x^2 - dy^2 = 1$. □

Example 18.5. Determine the integer solutions to the equation $x^2 - 54y^2 = 1$.
We have $\sqrt{54} = [7; \overline{2, 1, 6, 1, 2, 14}]$. We compute the convergents p_n/q_n to the continued fraction expansion of $\sqrt{54}$, using recurrence relations from class:

$$\frac{p_0}{q_0} = \frac{7}{1}$$
$$\frac{p_1}{q_1} = \frac{2 \cdot 7 + 1}{2},$$
$$\frac{p_2}{q_2} = \frac{1 \cdot 15 + 7}{1 \cdot 2 + 1},$$
$$\frac{p_3}{q_3} = \frac{6 \cdot 22 + 15}{6 \cdot 3 + 2},$$
$$\frac{p_4}{q_4} = \frac{1 \cdot 147 + 22}{1 \cdot 20 + 3},$$

$$\frac{p_5}{q_5} = \frac{2 \cdot 169 + 147}{2 \cdot 23 + 20} = \frac{485}{66},$$

so the fundamental solution of $x^2 - 54y^2 = 1$ is $(x, y) = (485, 66)$. Thus, the solutions (x, y) of $x^2 - 54y^2 = 1$ are given by $x + y\sqrt{54} = \pm(485 + 66\sqrt{54})^n$ $(n \in \mathbb{Z})$.

It is now easy to compute examples of further solutions. For example, since

$$(485 + 66\sqrt{54})^2 = 485^2 + 54 \cdot 66^2 + 2 \cdot 485 \cdot 66\sqrt{54},$$

we find that $(x, y) = (470449, 64020)$ is the second smallest solution of $x^2 - 54y^2 = 1$.

19. A sketch of binary quadratic forms (NON-EXAMINABLE)

A *binary quadratic form* is a homogeneous quadratic polynomial in two variables, which is to say, one of the shape $f(x, y) = ax^2 + bxy + cy^2$ (with fixed $a, b, c \in \mathbb{Z}$). The *discriminant* of this form is $d = b^2 - 4ac$.

Observe that

$$4af(x, y) = (2ax + by)^2 - (b^2 - 4ac)y^2 = (2ax + by)^2 - dy^2.$$

Thus, the binary form f is:

(a) *indefinite* (meaning that it takes both positive and negative values) when $d > 0$;
(b) *positive* (or *negative*) *definite* (meaning that it takes only positive (or only negative) values for $x, y \in \mathbb{Z}$) when $d < 0$;
(c) *positive* (or *negative*) *semi-definite* (meaning that it takes only non-negative (or only non-positive) values for $x, y \in \mathbb{Z}$) when $d = 0$.

Our goal in this section is to understand (in sketch form only) which integers, and in particular which prime numbers, are represented by binary quadratic forms.

We begin with a consideration of the discriminants d that can occur. Note that since

$$b^2 \equiv \begin{cases} 0 \pmod 4, & \text{when } 2|b, \\ 1 \pmod 4, & \text{when } 2 \nmid b, \end{cases}$$

then it follows that $d = b^2 - 4ac \equiv 0, 1 \pmod 4$. Moreover, both eventualities can occur. Indeed, the forms

$$\begin{cases} x^2 - \frac{1}{4}dy^2, & \text{when } d \equiv 0 \pmod 4, \\ x^2 + xy + \frac{1}{4}(1 - d)y^2, & \text{when } d \equiv 1 \pmod 4, \end{cases}$$

each have discriminant d in the respective cases, and are called the *principal* forms with discriminant d.

Next we consider the circumstances in which forms are morally speaking "the same", at least so far as the set of integers represented by the forms. For plainly $f(x, y)$ represents the same set of integers as does $f(\pm x, \pm y)$, for $x, y \in \mathbb{Z}$, and also the same set of integers as $f(x \pm y, y)$. This leads us to a consideration of invertible linear transformations between pairs of integers.

Definition 19.1. The group of 2×2 matrices with integral coefficients and determinant 1, denoted by $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, is called the *modular group*.

Definition 19.2. The quadratic forms

$$f(x,y)=ax^2+bxy+cy^2 \quad \text{and} \quad g(x,y)=Ax^2+Bxy+Cy^2$$

are *equivalent*, and we write $f\sim g$, if there exists $\gamma\in\Gamma$, say $\gamma=\begin{pmatrix} p & q \\ r & s\end{pmatrix}$, such that $g(x,y)=f(px+qy,rx+sy)$.

Theorem 19.3. *The relation $\sim$ between binary quadratic forms is an equivalence relation.*

Proof. This is a consequence of the observation that the identity matrix I belongs to Γ, that when $\gamma\in\Gamma$ then $\gamma^{-1}\in\Gamma$, and that when $\gamma_1,\gamma_2\in\Gamma$, then $\gamma_1\gamma_2\in\Gamma$. □

Note that, since the mapping $\gamma\in\Gamma$ is invertible with $\gamma^{-1}\in\Gamma$, it follows that when $f\sim g$, then the set of integers represented by $f(x,y)$ is identical to the set of integers represented by $g(x,y)$ (for $x,y\in\mathbb{Z}$). Moreover, we can view the quadratic form $f(x,y)=ax^2+bxy+cy^2$ in matrix form by observing that

$$ax^2+bxy+cy^2=(x,y)\begin{pmatrix} a & \frac{1}{2}b \\ \frac{1}{2}b & c\end{pmatrix}\begin{pmatrix} x \\ y\end{pmatrix}=\mathbf{x}^{\mathrm{T}}F\mathbf{x},$$

say, where $F=\begin{pmatrix} a & \frac{1}{2}b \\ \frac{1}{2}b & c\end{pmatrix}$. One then finds that in the situation described with $g\sim f$, one has

$$g(x,y)=(\gamma\mathbf{x})^{\mathrm{T}}F(\gamma\mathbf{x})=\mathbf{x}^{\mathrm{T}}(\gamma^{\mathrm{T}}F\gamma)\mathbf{x}.$$

This shows that

$$\mathrm{disc}(g)=-4\mathrm{det}(G)=-4\det(\gamma)^2\det(F)=-4\det(F)=\mathrm{disc}(f).$$

Thus, equivalent quadratic forms have the same discriminant.

Definition 19.4. We say that $f(x,y)$ *represents* $n\in\mathbb{Z}$ if there exists $(x_0,y_0)\in\mathbb{Z}$ with $f(x_0,y_0)=n$. This representation is *proper* if $(x_0,y_0)=1$.

Theorem 19.5. *When $f\sim g$, the (proper) representations of n by f are in one-to-one correspondence with the (proper) representations of n by g.*

Proof. Suppose that $f=\mathbf{x}^TF\mathbf{x}$ and $g=\mathbf{x}^TG\mathbf{x}$, and further that $\gamma\in\Gamma$ takes f to g. Then, if $f(x_0,y_0)=n$, one has $n=\mathbf{x}_0^TF\mathbf{x}_0$. But $\gamma\in\Gamma$ and $\gamma^{-1}\in\Gamma$, so

$$g(\gamma^{-1}\mathbf{x}_0)=(\gamma^{-1}\mathbf{x}_0)^TG(\gamma^{-1}\mathbf{x}_0)=\mathbf{x}_0^T(\gamma^{-1})^T\gamma^TF\gamma(\gamma^{-1}\mathbf{x}_0)=\mathbf{x}_0^TF\mathbf{x}_0=n,$$

so there is an injection from representations of n by f to those of n by g. Similarly, there is an injection from representations of n by g to those of n by f. We therefore conclude that the representations of n by f and those of n by g are in bijective correspondence.

In order to establish the claimed assertion concerning proper representations, it suffices to show that whenever

$$\gamma=\begin{pmatrix} p & q \\ r & s\end{pmatrix}\in\Gamma$$

and $(x,y)=1$, then with $\gamma\begin{pmatrix} x \\ y\end{pmatrix}=\begin{pmatrix} x' \\ y'\end{pmatrix}$, one has $(x',y')=1$. But with the latter hypotheses, one has that $(px+qy,rx+sy)$ divides

$$s(px+qy)-q(rx+sy)=(ps-qr)x=x,$$

and $(px+qy, rx+sy)$ divides

$$r(px+qy)-p(rx+sy)=-(ps-qr)y=-y.$$

Thus, we have $(px+qy,rx+sy)|(x,y)$, so that $(x',y')=(px+qy,rx+sy)=1$. □

Theorem 19.6. *Suppose that n and d are integers with $n\neq 0$. Then there exists a binary quadratic form of discriminant d which properly represents n if and only if the congruence $x^2\equiv d \pmod{4|n|}$ has a solution.*

Proof. We first suppose that the congruence $x^2\equiv d \pmod{4|n|}$ has a solution, say $b^2-d=4nc$, for some $b,c\in\mathbb{Z}$. Then the form

$$f(x,y)=nx^2+bxy+cy^2$$

has integral coefficients, discriminant $d=b^2-4nc$, and represents n properly, since $f(1,0)=n$. This establishes the reverse implication.

In the other direction, suppose that f has discriminant d and $n=f(p,r)$ for some $p,r\in\mathbb{Z}$ with $(p,r)=1$. By the Euclidean algorithm, there exist $q,s\in\mathbb{Z}$ with $ps-qr=1$, and then $f\sim f'$, with corresponding matrices $F'=\gamma^{\mathrm{T}}F\gamma$, and $\gamma=\begin{pmatrix} p & q\\ r & s\end{pmatrix}$. But then

$$f'(x,y)=a'x^2+b'xy+c'y^2=a(px+qy)^2+b(px+qy)(rx+sy)+c(rx+sy)^2,$$

where $a'=f(p,r)=n$. Moreover, since $f\sim f'$, we have $\mathrm{disc}(f')=\mathrm{disc}(f)$, and thus $b'^2-4nc'=d$. In particular, the congruence $x^2\equiv d \pmod{4|n|}$ has solution $x=b'$. □

Observe that if there is *just one* equivalence class of binary forms with discriminant d, then Theorem 19.6 can be applied to understand which integers are represented by any fixed binary quadratic form of discriminant d. This is because, in such circumstances, any fixed form of discriminant d is equivalent to any other form of discriminant d.

We now come to the big idea of Gauss in this context, that of classifying the individual binary quadratic forms representing each equivalence class of forms.

Definition 19.7. Let $f(x,y)=ax^2+bxy+cy^2$ be a binary quadratic form whose discriminant d is not a perfect square. We say that f is *reduced* if either

$$-|a|<b\leqslant |a|<|c|,$$

or

$$0\leqslant b\leqslant |a|=|c|.$$

Note that when d is a perfect square, then these conditions are replaced by requiring that

$$c=0 \quad\text{and}\quad 0\leqslant a<|b|,\quad \text{when } d>0,$$

and

$$a\geqslant 0 \quad\text{and}\quad b=c=0,\quad \text{when } d=0.$$

Theorem 19.8. *If f is a binary quadratic form with discriminant d not equal to a perfect square, then f is equivalent to a reduced binary quadratic form.*

Definition 19.9. We write $H(d)$ for the number of equivalence classes of binary quadratic forms of discriminant d with leading coefficient positive. This quantity $H(d)$ is known as the *class number of d*.

One can check that $H(d) \leqslant 2d$ when $d > 0$, and that $H(d) \leqslant \frac{8}{3}|d|$ when $d < 0$. A highlight of the mid-twentieth century was the following result.

Theorem 19.10 (Baker, 1966; Stark, Heegner). *When $d < 0$, one has $H(d) = 1$ only when $d = -3, -4, -7, -8, -11, -19, -43, -67, -163$.*

Example 19.11. We determine the prime numbers represented by the positive definite binary form x^2+y^2. This is of course a familiar example that we have previously analysed, but serves to illustrate the ideas of this section. Here we have a positive definite binary form of discriminant $d = -4$. We observe that if a reduced positive definite binary quadratic form $f(x,y) = ax^2 + bxy + cy^2$ has discriminant -4, then we have $a > 0$, $c > 0$ and $b^2 - 4ac = -4$, and hence $4 = 4ac - b^2 \geqslant 3ac$. Thus $3ac \leqslant 4$, which ensures that $a = c = 1$. Finally, this shows that $4 = 4 - b^2$, whence $b = 0$. So the only reduced binary quadratic form of discriminant -4 is $f(x,y) = x^2 + y^2$, whence $H(-4) = 1$, and in particular all positive definite binary quadratic forms of discriminant -4 with positive leading coefficient are equivalent. We therefore deduce from Theorem 19.6 that p is represented by $x^2 + y^2$ if and only if $x^2 \equiv -4 \pmod{4p}$ is soluble. When $p = 2$, we check that $x^2 \equiv -4 \pmod 8$ has solution $x = 2$, and when p is odd we instead determine the conditions in which

$$1 = \left(\frac{-4}{p}\right) = \left(\frac{-1}{p}\right) = (-1)^{(p-1)/2},$$

a condition that holds if and only if $p \equiv 1 \pmod 4$. Hence we see that the prime p is represented by $x^2 + y^2$ if and only if either $p = 2$ or $p \equiv 1 \pmod 4$.

Example 19.12. We determine the prime numbers represented by the positive definite binary quadratic forms of disriminant -11. We observe that if a reduced positive definite binary quadratic form $f(x,y) = ax^2+bxy+cy^2$ has discriminant -11, then we have $a > 0$, $c > 0$ and $b^2 - 4ac = -11$, and hence $11 = 4ac - b^2 \geqslant 3ac$ as well as $11 \leqslant 4ac$. Thus, one has $11/4 \leqslant ac \leqslant 11/3$, which ensures that $ac = 3$, and hence $b^2 = 4ac - 11 = 1$. So the only reduced positive definite binary quadratic form of discriminant -11 is the principal form $f(x,y) = x^2 + xy + 3y^2$ with $b = 1$, $a = 1$ and $c = 3$. Hence $H(-4) = 1$, and in particular all positive definite binary quadratic forms of discriminant -11 are equivalent. We therefore deduce from Theorem 19.6 that p is represented by $x^2+xy+3y^2$ if and only if $x^2 \equiv -11 \pmod{4p}$ is soluble. Since $x^2 \equiv -3 \pmod 8$ is not soluble, it follows that 2 is not represented. When p is odd, we have conditions to check modulo 4 and modulo p. The first condition is checked by noting that $x^2 \equiv -11 \pmod 4$ has solution $x = 1$. For the second to be satisfied, when $p \neq 11$ we require

$$1 = \left(\frac{-11}{p}\right) = (-1)^{(p-1)/2}\left(\frac{11}{p}\right) = (-1)^{(p-1)/2}(-1)^{(11-1)(p-1)/4}\left(\frac{p}{11}\right) = \left(\frac{p}{11}\right).$$

This condition is satisfied if and only if p is a square modulo 11, so p is congruent to one of $1, 4, 9, 16, 25$ modulo 11. When $p = 11$, meanwhile, the congruence $x^2 \equiv -11 \pmod p$ is trivially soluble. We therefore conclude that the prime p is represented by $x^2+xy+3y^2$ if and only if $p = 11$ or $p \equiv 1, 3, 4, 5, 9 \pmod{11}$.

Example 19.13. Determine the equivalence classes of positive definite binary quadratic forms of discriminant -36, and deduce which prime numbers are represented by these equivalence classes.

We observe that if a reduced positive definite binary quadratic form

$$f(x,y)=ax^2+bxy+cy^2$$

has discriminant -36, then we have $a>0$, $c>0$ and $b^2-4ac=-36$, and hence $36=4ac-b^2\geqslant 3ac$ as well as $36\leqslant 4ac$. Thus, one has $9\leqslant ac\leqslant 12$, as well as $1\leqslant a\leqslant c$. We divide into cases:
(i) $a=3$ and $c=4$. Then $4\cdot 3\cdot 4-b^2=36$, whence $b^2=12$, which is not possible.
(ii) $a=c=3$. Then $4\cdot 3\cdot 3-b^2=36$, whence $b^2=0$. This yields the reduced form $f(x,y)=3x^2+3y^2$.
(iii) $a=2$ and $c=6$. Then $4\cdot 2\cdot 6-b^2=36$, whence $b^2=12$, which is not possible.
(iv) $a=2$ and $c=5$. Then $4\cdot 2\cdot 5-b^2=36$, whence $b^2=4$. Thus $b=\pm 2$, but since $b>-a$, it follows that $(a,b,c)=(2,2,5)$, and we obtain the reduced form $f(x,y)=2x^2+2xy+5y^2$.
(v) $a=1$ and $c=12$. Then $4\cdot 1\cdot 12-b^2=36$, whence $b^2=12$, which is not possible.
(vi) $a=1$ and $c=11$. Then $4\cdot 1\cdot 11-b^2=36$, whence $b^2=8$, which is not possible.
(vii) $a=1$ and $c=10$. Then $4\cdot 1\cdot 10-b^2=36$, whence $b^2=4$. We then have $b=\pm 2$, and yet we require $-a<b\leqslant a$, so this is not possible.
(viii) $a=1$ and $c=9$. Then $4\cdot 1\cdot 9-b^2=36$, whence $b^2=0$. This yields the reduced form $f(x,y)=x^2+9y^2$.

So the only reduced positive definite binary quadratic forms of discriminant -36 are

$$3(x^2+y^2),\quad 2x^2+2xy+5y^2,\quad x^2+9y^2.$$

Hence $H(-36)=3$.

From Theorem 19.6 we find that p is represented by some one of these forms if and only if $x^2\equiv -36 \pmod{4p}$ is soluble. Since $x^2\equiv 4\pmod 8$ is always soluble, the prime number 2 is represented, and it is apparent that the only form representing 2 is $2x^2+2xy+5y^2$. When $p\geqslant 3$, meanwhile, the congruence $x^2\equiv -36\pmod 4$ is trivially soluble. It remains to examine the congruence modulo p. Here, when $p=3$, we find that the congruence is again trivially soluble. While $3(x^2+y^2)$ plainly represents 3, the other two forms do not. To see this, observe that the smallest two values represented by x^2+9y^2 are 1 and 4, and likewise the smallest two values represented by $2(2x^2+2xy+5y^2)=(2x+y)^2+9y^2$ are 4 and 10. Thus 3, and respectively, 6 are not represented, and we deduce that neither x^2+9y^2 nor $2x^2+2xy+5y^2$ represent 3. Finally, when $p>3$, we have $x^2\equiv -36\pmod p$ soluble if and only if

$$1=\left(\frac{-36}{p}\right)=(-1)^{(p-1)/2}.$$

Then we must have $p\equiv 1\pmod 4$. We observe that when $p>3$ and $p=a^2+9b^2$, then $3\nmid a$, whence $p\equiv 1\pmod 3$. Also, when $p>3$ and $p=2a^2+2ab+5b^2$, then $2p=(2a+b)^2+9b^2$, so that $3\nmid(2a+b)$, whence $2p\equiv 1\pmod 3$, and hence $p\equiv 2\pmod 3$. This division between congruence classes shows that p is represented by x^2+9y^2 whenever $p\equiv 1\pmod 4$ and $p\equiv 1\pmod 3$, and p is represented by $2x^2+2xy+5y^2$ whenever $p\equiv 1\pmod 4$ and $p\equiv 2\pmod 3$.

In conclusion, the form $3(x^2+y^2)$ represents the prime 3 only, the form x^2+9y^2 represents the primes p with $p\equiv 1\pmod{12}$, and the form $2x^2+2xy+5y^2$ represents the prime 2 and the primes p with $p\equiv 5\pmod{12}$.

Department of Mathematics, Purdue University, N. University Street, West Lafayette,
West Lafayette, IN 47907-2067, USA

Email address: twooley@purdue.edu