MA 7: NUMBER THEORY FOR BEGINNERS

Introduction

These are lecture notes for Ma 7/107 "Number Theory for Beginners" at Caltech. They are based on the hand-written notes of Serin Hong, who taught this course in 2015-2016.

Contents

Notation. Some standard notation:

$$\begin{aligned} \mathbb{N} &: \text{ the natural numbers } \{1,2,3,\cdots\} \\ \mathbb{Z} &: \text{ the integers } \{\cdots,-2,-1,0,1,2,\cdots\}. \\ \mathbb{Z}_{\geq 0} &: \text{the non-negative integers } = \mathbb{N}\cup\{0\} \\ \mathbb{Q} &: \text{the rational numbers} \end{aligned}$$

1. Division, GCD, Euclidean Algorithm

Given two integers a and b, among the four quantities

$$a + b, \quad a - b, \quad a \cdot b, \quad a/b,$$

it's the last one that is not guaranteed to be an integer. Investigating which integers are divisible by which others is the very beginning of number theory.

Definition 1.1. Suppose that for $a, b \in \mathbb{Z}$, there exists $q \in \mathbb{Z}$ with $b = qa$. We express this fact in equivalent ways by saying:

- b is *divisible* by a
- b is a *multiple* of a
- a is a *divisor* of b
- a *divides* b

and we write: $a|b$.

Proposition 1.2. *The following propositions hold:*

(1) $\forall a \in \mathbb{Z} \ \ a|0, \ 1|a.$
(2) $a|b, \ c \in \mathbb{Z} \Longrightarrow a|bc$
(3) $a|b, \ b|c \Longrightarrow a|c$
(4) $a|b, \ a|c \Longrightarrow a|b \pm c$

Proof. These are intuitive and easy to verify formally. For example if $b = q_1 a$ and $c = q_2 b$ then

$$c = q_2 b = q_2(q_1 a) = (q_2 q_1)a,$$

which proves (3).

Note that the key fact used is the commutativity of multiplication in the step: $q_2(q_1 a) = (q_2 q_1)a$. We leave the other parts as exercise. Which basic properties of numbers are needed to prove them? □

Corollary 1.3. *Suppose that $a|b$ and $a|c$. Then $a|mb + nc$ for any $m, n \in \mathbb{Z}$.*

Proof. Combine property (2) and (4). □

Definition 1.4. Suppose $a, b \in \mathbb{Z}$, at least one of them non-zero. The **greatest common divisor** (**gcd**) of a and b is the largest integer d such that $d|a$ and $d|b$. We write: $(a, b) = d$.

Note that $(a, b) > 0$ always, unless $a = b = 0$. In that case $(0, 0) = 0$ by convention.

Example: The (positive) divisors of 60 are:

$$1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60$$

The positive divisors of 70 are

$$1, 2, 5, 7, 10, 14, 35, 70.$$

The common divisors have been colored. The largest among them is $10 = (70, 60)$.

Definition 1.5. The **least common multiple** (**lcm**) of a and b is the smallest *positive* integer l such that $a|l$ and $b|l$.

The positive multiples of 60:

$$60, 120, 180, 240, 300, 360, 420, 480, 540, \cdots$$

and the positive multiples of 70 :

$$70, 140, 210, 280, 350, 420, 490, 560, \cdots$$

The first common number is 420, so that's the lcm of 60 and 70.

As we will see, there are better ways to compute gcd and lcm than making long lists.

Definition 1.6. Two integers a, b are said to be **relatively prime** if $(a, b) = 1$.

Euclidean Algorithm.

Proposition 1.7. *Given any two integers a and b, with $b \neq 0$, there exist unique non-negative integers q and r such that:*

$$a = qb + r, \quad 0 \leq r < b.$$

Proof. First we prove the *existence* of q and r:

If $b|a$, i.e. $a = qb$ for some q, then $r = 0$ and there is nothing to prove. If $b \nmid a$, then a falls strictly between two consecutive multiples of b, say qb and $(q+1)b$. Then

$$qb < a < qb + b \Longrightarrow 0 < a - qb < b$$

so for $r = a - qb$ we have $a = qb + r$ and $0 \leq r < b$.

Now to prove *uniqueness* of q and r. Suppose $a = bq + r = bq' + r'$ with $0 \leq r, r' \leq b$. Then

$$r' + q'b = r + qb \Longrightarrow r - r' = b(q' - q).$$

Then r, r' are non-negative integers strictly smaller than b, whose difference is a multiple of b. This is only possible if that multiple is 0, i.e. $r = r'$. Then $b(q' - q) = 0$ and so $q = q'$ since $b \neq 0$.

To be careful we also do this a bit more formally. If $q \neq q'$, then without loss of generality we can assume $q' > q$. Then

$$q' - q \geq 1 \Longrightarrow b(q' - q) \geq b \Longrightarrow r = r' + b(q' - q) \geq r' + b \geq b.$$

But $r \geq b$ contradicts $0 \leq r < b$, so the assumption that $q \neq q'$ must be false. Then from $q = q'$ we obtain $r = a - q'b = a - qb = r'$. This proves uniqueness of q and r. □

Of course q and r are the familiar *quotient* and *remainder* of the division of a by b.

The following observation was used by Euclid, around 300 BC, in order to calculate the gcd.

Lemma 1.8. *Suppose $a = bq + r$. Then $(a, b) = (b, r)$.*

Proof. Let d be a divisor of b. Then $d|bq$ by property (2) of Proposition 1.2(2). If $d|r$ we have $d|bq + r$ by property (4), i.e. $d|a$. Conversely, if $d|a$ we have

$$d|(a - bq) = bq + r - bq = r,$$

so $d|a$ iff $d|r$. This shows the common divisors of b and r are exactly the same as the common divisors of a and b. Then the greatest among them is both (a, b) and (b, r). □

Corollary 1.9. *For all $a, b \in \mathbb{Z}$, $(a, b) = (a, a - b)$.*

The lemma allows us to compute the gcd of any two integers a and b using the following recursive sequence of steps, called the **Euclidean algorithm**:

(0) If $b < 0$ replace it with $-b$ and note that it doesn't change (a, b).
(1) Write $a = bq + r$, for $0 \leq r < b$. Note that $(a, b) = (b, r)$ by Lemma 1.8.
(2) If $r = 0$, then $(a, b) = (b, 0) = b$.
(3) If $r \neq 0$, return to step 1 with (a, b) replaced by (b, r).

Each time the algorithm reaches step (3), b is replaced by a strictly smaller non-negative number. Since that can't happen indefinitely, the process must terminate at some point, i.e. eventually $r = 0$ in step (1) giving the answer in step (2).

Example: Finding the gcd of -1740 and 522:

$$\begin{aligned} -1740 &= 522 \cdot (-4) + 348 \\ 522 &= 348 \cdot 1 + 174 \\ 348 &= 174 \cdot 2 + 0 \end{aligned}$$

So we have $(-1740, 522) = (522, 348) = (348, 174) = (174, 0) = 174$.

As another example, we can apply what we've learned so far to the *Fibonacci numbers* F_n. This is the sequence of numbers specified by

$$F_0 = 0,\ F_1 = 1, \quad F_{n+1} = F_n + F_{n-1} \ \ n > 0.$$

The first few are:

$$0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ 55,\ 89,\ \cdots$$

Theorem 1.10. *All consecutive pairs of Fibonacci numbers are relatively prime.*

Proof. The proof is an example of mathematical induction. Let P_n denote the statement:

$$F_n \text{ and } F_{n+1} \text{ are relatively prime.}$$

We want to prove P_n is true for all $n \geq 0$.

Base case: For $n = 0$, we have $(F_0, F_1) = (0, 1) = 1$. This shows P_0 is true.

Induction step: Assume P_k is true, i.e. that $(F_k, F_{k-1}) = 1$. Then we have

$$(F_{k+1}, F_k) = (F_k + F_{k-1}, F_k) = (F_k, F_{k-1}) \quad \text{(by Corollary 1.9).}$$

Then

$$P_k \text{ is true } \Longrightarrow (F_{k+1}, F_k) = 1 \Longrightarrow (F_k, F_{k-1}) = 1 \Longrightarrow P_{k+1} \text{ is true .}$$

The base case and the induction step together imply, by the *principle of mathematical induction*, that P_n is true for *all* $n \geq 0$. □

Using the gcd we can determine whether or not simple equations of the form $ax + by = c$, $a, b, c \in \mathbb{Z}$ have integer solutions.

Theorem 1.11. *Suppose $a, b, c \in \mathbb{Z}$, $a, b \neq 0$. The equation $ax + by = c$ has a solution with $x, y \in \mathbb{Z}$ if and only if $(a, b)|c$.*

Proof. Let $g = (a, b)$. If $ax + by = c$ with $x, y \in \mathbb{Z}$, then from $g|a$ and $g|b$ we have $g|ax + by$ (by Corollary 1.3). This shows the condition $(a, b)|c$ is *necessary* for the existence of solutions $x, y \in \mathbb{Z}$.

Now consider the Euclidean algorithm applied to a and b:

$$\begin{aligned} a &= bq + r, & 0 &\leq r < b \\ b &= rq_1 + r_1, & 0 &\leq r_1 < r \\ r &= r_1 q_2 + r_2, & 0 &\leq r_2 < r_1 \\ &\vdots & & \\ r_{n-1} &= r_n \cdot q_{n+1} + 0 & & \end{aligned}$$

which ends for some n, and $r_n = g$ as we have seen. If we let $r_0 = r$, $r_{-1} = b$ and $r_{-2} = a$, The equations all have the form $r_{k-1} = r_k q_{k+1} + r_{k+1}$, for $-1 \leq k \leq n$.

We will prove by induction that each $ax + by = r_k$ has a solution. Let P_k denote the statement:

$$ax + by = r_k \text{ has a solution.}$$

Base cases: Note

$$ax + by = r$$

has a solution $x = 1$, $y = -q$, since $a = bq + r$. This shows P_0 is true. Also

$$ax + by = b$$

has the solution $x = 1$, $y = 0$, so P_{-1} is also true.

Induction step: Assume that for some positive integer m, P_k is true for all $-1 \leq k < m$. In other words that there exist integers x_k, y_k with $ax_k + by_k = r_k$. In particular, there exist integers x_{m-1}, y_{m-1}, x_{m-2}, y_{m-2} such that

$$r_{m-1} = ax_{m-1} + by_{m-1}, \quad r_{m-2} = ax_{m-2} + by_{m-2}.$$

Then

$$r_{m-2} = q_{m-1}r_{m-1} + r_m \Longrightarrow ax_{m-2} + by_{m-2} = q_{m-1}(ax_{m-1} + by_{m-1}) + r_m$$

so that

$$r_m = a(x_{m-2} - a_{m-1}x_{m-1}) + b(y_{m-2} - q_{m-1}y_{m-1}).$$

This shows P_m is true.

By induction, P_n is also true, so that $g = ax + by$ has integer solutions $x = x_n, y = y_n$.

Now let us return to the equation

$$ax + by = c$$

and suppose $g|c$, so $c = gc_0$. Then from $ax_n + by_n = g$ we obtain

$$a(x_n c_0) + b(y_n c_0) = gc_0 = c.$$

This shows the condition $g|c$ is also *sufficient* for $ax + by = c$ to have a solution $x, y \in \mathbb{Z}$. □

Corollary 1.12. *$ax+by = 1$ has a solution with $x, y \in \mathbb{Z}$ if and only if a and b are relatively prime.*

Corollary 1.13. *Suppose for $n, a, b \in \mathbb{Z}$ that $n|ab$. If $(a, n) = 1$, then $n|b$.*

Proof. If $(a, n) = 1$, then $ax_0 + ny_0 = 1$ for some $x_0, y_0 \in \mathbb{Z}$. From $n|ab$ we have $ab = mn$ for some $m \in \mathbb{Z}$. Then

$$b = bax_0 + bny_0 = n(mx_0 + by_0) \Longrightarrow n|b.$$

□

2. Linear Systems in Two variables, Prime Factorizations

For $a, b, c \in \mathbb{Z}$, we have seen how to tell whether

$$ax + by = c$$

has a solution with $x, y \in \mathbb{Z}$.

Example: Suppose we are given

$$1342x + 154y = 198 \tag{1}$$

We first find $(1342, 154)$:

$$\begin{aligned} 1342 &= 154 \cdot 8 + 110 \\ 154 &= 110 \cdot 1 + 44 \\ 110 &= 44 \cdot 2 + 22 \\ 44 &= 22 \cdot 2 + 0 \end{aligned}$$

So $(1342, 154) = 22$. We find that $22|198$ so the equation above *does* have solutions $x, y \in \mathbb{Z}$. We have

$$\begin{aligned}22 &= 110 - 44 \cdot 2\\ &= (1342 - 154 \cdot 8) - (154 - 110 \cdot 1) \cdot 2\\ &= 1342 - 154 \cdot 8 - 154 \cdot 2 + (1342 - 154 \cdot 8) \cdot 2\\ &= 1342 \cdot 3 - 154 \cdot 26\end{aligned}$$

Now $198 = 9 \cdot 22$, so from

$$1342 \cdot 3 + 154 \cdot (-26) = 22$$

multiplying by 9 we get

$$1342 \cdot 27 + 154 \cdot (-234) = 198.$$

Remark 2.1. The solution of such an equation is never unique. In general, if (x_0, y_0) is a solution of $ax+by = c$, then (x_0+b, y_0-a) is another solution. To be thorough, we would like to characterize *all* the solutions.

Remark 2.2. Note that we could have simplified (1) by dividing by 22, obtaining

$$61x + 7y = 9. \tag{2}$$

We then have $(61, 7) = 1$.

In general we have:

Lemma 2.3. *Suppose $g = (a, b) > 0$ for $a, b \in \mathbb{Z}$. Then*

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

Proof. If d is a positive common divisor of a/g and b/g, dg is a common divisor of a and b. But $dg \geq g$, so $gd = g$, $d = 1$. □

Theorem 2.4. *Suppose $a, b, c \in \mathbb{Z}$ are non-zero, and $g|c$, where $g = (a, b)$. If (x_0, y_0) is a particular solution of $ax + by = c$, the general solution parametrized by $t \in \mathbb{Z}$ is*

$$\begin{cases} x = x_0 - \frac{b}{g}t \\ \\ y = y_0 + \frac{a}{g}t. \end{cases}$$

Proof. First we assume $(a, b) = 1$.

Suppose (x_1, y_1) is a different solutions of $ax + by = c$. Then

$$ax_0 + by_0 = ax_1 + by_1 \implies a(x_0 - x_1) = b(y_1 - y_0).$$

We then have $a|b(y_1 - y_0)$. But $(a, b) = 1$ so by Lemma 1.13 we have $a|y_1 - y_0$. Then $y_0 - y_1 = ta$ for some $t \in \mathbb{Z}$. Then we have $a(x_0 - x_1) = abt \implies (x_0 - x_1) = bt$. In other words,

$$x_1 = x_0 - bt$$
$$y_1 = y_0 + at.$$

Conversely, as we noted before, for any $t \in \mathbb{Z}$ the above expressions give a new solution (x_1, y_1).

Now in general if $(a, b) = g$ and $g|c$, we can write $a = a'g$, $b = b'g$, $c = c'g$. Then $ax + by = c$ if and only if $a'x + b'y = c'$. By Lemma 2.3 we have $(a', b') = 1$, so as we have just seen the general solution to $ax + by = c$ is

$$x = x_0 - b't$$
$$y = y_0 + a't.$$

□

We have completely solved the problem of finding solutions to $ax + by = c$ for $a, b, c \in \mathbb{Z}$. The criterion for whether or not any solution exists depended on the $g = (a, b)$. This already gives an indication that the divisors of integers play an important role in number theory.

Definition 2.5. A natural number n is called **prime** if it has exactly two positive divisors. It is called **composite** if it has more.

The definition is formulated in such a way so as to exclude the number 1 from being either prime or composite. This is deliberate, owing to the exceptional status of 1. Every $n > 1$ has at least two distinct divisors, 1 and itself, so all $n > 1$ are either prime or composite. The first few primes are

$$2,\ 3,\ 5,\ 7,\ 11,\ 13,\ 17,\ 19,\ 23,\ 29, 31,\ \cdots$$

Lemma 2.6. *Suppose p is a prime number, $a, b \in \mathbb{Z}$, and $p|ab$. Then either $p|a$ or $p|b$.*

Proof. Suppose $p \not| a$. Then $(a, p) = 1$, since p has no other factor to share with a other than 1. By Corollary 1.13 we have $p|b$. □

Corollary 2.7. *If $p|a_1 \cdots a_n$, with $a_i \in \mathbb{N}$ and p prime, then $p|a_i$ for some $i \in \{1, \cdots, n\}$.*

Proof. Left as an exercise in induction. □

Definition 2.8. A **prime factorization** for a natural number $n > 1$, is an expression of the form

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r},$$

where $p_1, \cdots, p_r$ are distinct (positive) primes, and $e_1, \cdots, e_r \in \mathbb{N}$.

We'll soon prove that every natural number n has a prime factorization. For instance

$$12345 = 3 \cdot 5 \cdot 823.$$

Of course there are often repeated factors, as in

$$23456 = 2^5 \cdot 733.$$

The prime numbers are then the indivisible atoms of the natural numbers.

Theorem 2.9 (Fundamental Theorem of Arithmetic)**.** *Every integer $n > 1$ has a unique prime factorization, up to rearrangement of the prime factors.*

Proof. First we prove the *existence* of prime factorizations, by induction.

Base case: $n = 2$. Since 2 is itself prime, the trivial identity $2 = 2$ is a prime factorization for 2.

Induction step: Assume that $n > 2$, and that any k with $2 \leq k < n$ has a prime factorization (induction hypothesis).

If n is prime, there's nothing to show, since again $n = n$ is a prime factorization of n. Otherwise $n = ab$, where $1 < a, b < n$. By the induction hypothesis

$$a = p_1^{e_1} \cdots p_r^{e_r}, \quad b = q_1^{f_1} \cdots q_s^{f_s}$$

where $\{p_1, \cdots, p_r\}$ and $\{q_1, \cdots, q_s\}$ are each a set of distinct primes, and e_i, $f_i \in \mathbb{N}$. Then we have

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r} q_1^{f_1} q_2^{f_2} \cdots q_s^{f_s}.$$

To obtain a prime factorization of n, the primes have to be distinct. Starting with the above, if any $p_i = q_j$ we can combine the corresponding factors $p_i^{e_i} \cdot q_j^{f_j}$ and write it as $p_i^{e_i + f_j}$. Continuing to combine factors in this way, each time the number of primes in the expression decreases. Since this number can't be zero, at some point all the primes will be distinct, so that we obtain a prime factorization

$$n = P_1^{d_1} \cdots P_r^{d_r}.$$

Now we show *uniqueness* of prime factorizations.

Suppose that

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r} = q_1^{e'_1} q_2^{e'_2} \cdots q_s^{e'_s},$$

are two prime factorizations of n. Let p be one of the primes p_i. By Corollary 2.7 $p|q_j$ for some j, hence $p = q_j$. Then each prime p_i is among the primes $q_1, \cdots, q_s$, so that $\{p_1, \cdots, p_r\} \subseteq \{q_1, \cdots, q_s\}$. By the same argument, $\{q_1, \cdots, q_s\} \subseteq \{p_1, \cdots, p_r\}$, so that the set $\{p_1, \cdots p_r\}$ must coincide with $\{q_1, \cdots q_s\}$. In particular $r = s$. By rearranging indices if necessary we can assume $q_i = p_i$, $1 \leq i \leq r$. We must then show $e_i = e'_i$. We proceed by induction on r, the number of prime factors.

For the base case $r = 1$, and $n = p^e = p^{e'}$. If $e \neq e'$, wolog $e > e'$, so that $p^e = p^{e'} \cdot p^{e-e'} > p^{e'}$, which contradicts $p^e = p^{e'}$, so that $e = e'$.

Now assume that $r > 1$ and that all natural numbers n with $r - 1$ distinct prime factors have a unique prime factorization (up to rearrangement). Suppose

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r} = p_1^{e'_1} p_2^{e'_2} \cdots p_r^{e'_r}$$

are two prime factorizations for n. Wolog assume $e_1 \leq e'_1$. Then

$$m = np^{-e_1} = p_2^{e_2} \cdots p_r^{e_r} = p_1^{e'_1 - e_1} p_2^{e'_2} \cdots p_r^{e'_r}.$$

Since $\{p_1, \cdots, p_r\}$ are distinct, $p_1 \notin \{p_2, \cdots, p_r\}$, so that $p_1 \not| m$ using $m = p_2^{e_2} \cdots p_r^{e_r}$. But then from $p_1^{e'_1 - e_1} p_2^{e'_2} \cdots p_r^{e'_r}$ we must have $e'_1 - e_1 = 0$. Then we obtain

$$m = p_2^{e_2} \cdots p_r^{e_r} = p_2^{e'_2} \cdots p_r^{e'_r}.$$

By the induction hypothesis, these two prime factorization are the same, therefore $e_i = e'_i$ for all $i > 2$. Plus $e_1 = e'_1$ this shows the two prime factorizations of n coincide. We have shown that all numbers n with r distinct prime factors have prime factorizations, assuming all those with $r - 1$ prime factors do.

By induction, all natural numbers $n > 1$ have a unique prime factorization. □

By convention, we may consider $1 = 1$ to be the prime factorization of 1.

Corollary 2.10. *Suppose that $n \in \mathbb{N}$ has prime factorization*

$$n = p_1^{e_1} \cdots p_r^{e_r}.$$

Then for $d \in \mathbb{N}$, we have $d|n$ if and only if

$$d = p_1^{e'_1} \cdots p_r^{e'_r}$$

with each $0 \leq e'_i \leq e_i$.

Proof. If d has the form indicated, then $d|n$ since

$$n = p_1^{e_1 - e'_1} \cdots p_r^{e_r - e'_r}.$$

Suppose now that $n = dm$ and that

$$d = q_1^{f_1} \cdots q_s^{f_s}, \quad m = q_1'^{f'_1} \cdots q_t'^{f'_t}$$

are the prime factorizations of d and m. Then

$$n = q_1^{f_1} \cdots q_s^{f_s} \cdot q_1'^{f'_1} \cdots q_t'^{f'_t}.$$

If we add the primes $\{q'_j\}$ to the set $\{q_i\}$, we can write $\{q_1, \cdots q_{s'}\} = \{q_1, \cdots q_s\} \cup \{q'_1, \cdots q'_t\}$, where $s' \geq s$. Then

$$n = q_1^{g_1}, \cdots q_{s'}^{g_{s'}}.$$

where for $i \leq s$, $f_i \leq g_i$. But by the uniqueness of prime factorization for n, we have $r = s'$, and $q_i = p_i$ up to rearrangement. In particular, each prime $q_i = p_j$ for some $j \leq r$, and $f_i \leq g_i = e_j$. Letting $e'_j = f_i$ for such primes p_j and $e'_j = 0$ otherwise, we obtain

$$d = p_1^{e'_1} \cdots p_r^{e'_r}$$

as expected. □

3. Fundamental Theorem of Arithmetic

Suppose that for $n \in \mathbb{N}$, we have a prime factorization

$$n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}.$$

We saw that positive divisors d of n are all the numbers of the form

$$d = p_1^{e'_1} p_2^{e'_2} \cdots p_r e'_r.$$

with $0 \leq e'_i \leq e_i$. For each $i \in \{1, \cdots, r\}$, the numbers e'_i can take any of the $e_i + 1$ values $\{0, 1, \cdots, e_i\}$. Since prime factorization is unique, the total number of positive divisors of n is

$$(e'_1 + 1)(e'_2 + 1) \cdots (e'_r + 1).$$

Example In the first lecture, we made a list of positive divisors of 60:

$$1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60$$

We have $60 = 2^2 \cdot 3 \cdot 5$, so 60 has $(2+1) \cdot (1+1) \cdot (1+1) = 12$ positive divisors. This shows the list is complete.

The number of positive divisors of n is denoted $\sigma_0(n)$. It is the case $k = 0$ of the following.

Definition 3.1. For $n \in \mathbb{N}$, $k \in \mathbb{Z}_{\geq 0}$ we put

$$\sigma_k(n) = \sum_{\substack{d>0 \\ d|n}} d^k.$$

For instance $\sigma_1(60)$ is the *sum* of the positive divisors of 60. Rather than add the list of 12 numbers above, we can use the following formula.

Proposition 3.2. *Suppose $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ is the prime factorization of n. For $k \in \mathbb{Z}_{\geq 0}$ we have*

$$\sigma_k(n) = \prod_{i=1}^{r} \frac{p_i^{(e_i+1)k} - 1}{p_i^k - 1} = \frac{p_1^{(e_1+1)k} - 1}{p_1^k - 1} \cdot \frac{p_2^{(e_2+1)k} - 1}{p_2^k - 1} \cdots \frac{p_r^{(e_r+1)k-1}}{p_r^k - 1}.$$

Proof. Since $\frac{p^{k(e+1)-1}}{p^k-1} = 1 + p^k + p^{2k} + \cdots + p^{ek}$, the product above is

$$(1 + p_1^k + \cdots + p_1^{ke_1})(1 + p_2^k + \cdots + p_2^{ke_2}) \cdots (1 + p_r^k + \cdots + p_r^{ke_r}).$$

If we expand this sum, it will be a sum of numbers of the form

$$p_1^{ke'_1} \cdot p_2^{ke'_2} \cdots p_r^{ke'_r},$$

for $0 \leq e'_i \leq e_i$. This is the exactly d^k for $d = p_1^{e'_1} \cdot p_2^{e'_2} \cdots p_r^{e'_r}$. □

Example: In the special case $k = 1$, we have

$$\sigma_1(n) = \frac{p_1^{e_1+1} - 1}{p_1 - 1} \cdots \frac{p_r^{e_r+1} - 1}{p_r - 1}.$$

In particular, for $60 = 2^2 \cdot 3 \cdot 5$,

$$\sigma_1(60) = \frac{2^3-1}{2-1} \cdot \frac{3^2-1}{3-1} \cdot \frac{5^2-1}{5-1} = 7 \cdot 4 \cdot 6 = 168.$$

Let $S = \{d \in \mathbb{N} : d|n\}$ be the set of positive divisors of n. We have a function

$$f : S \longrightarrow S, \quad f(d) = \frac{n}{d}.$$

The function f a bijection, and it is its own inverse: $f(f(d)) = d$. (Such a function is called an *involution.*) Note that n is a square if and only if f has a fixed point. In other words, if and only if there exists $m \in S$ such that $f(m) = m$. Furthermore, there is at most one such fixed point.

Proposition 3.3. *For $n \in \mathbb{N}$, $\sigma_0(n)$ is odd if and only if n is a square.*

Proof. If f has no fixed point, we can partition S into distinct pairs $\{d, f(d)\}$, so in that case $|S| = \sigma_0(n)$ must be even.

If f does have a fixed point, i.e. if $n = m^2$, we can partition $S - \{m\}$ into pairs $\{d, f(d)\}$. So $|S - \{m\}| = \sigma_0(n) - 1$ must be even. □

The gcd and lcm via prime factorizations. Suppose $a, b \in \mathbb{N}$ and that $p_1, \cdots, p_r$ are all the distinct prime factors of a and b combined. Then

$$a = p_1^{e_1} \cdots e_r^{e_r}, \quad b = p_1^{e'_1} \cdots p_r^{e'_r}$$

where now some e_i, e'_i are allowed to be zero.

If d is a divisor of a, then its prime factors are among the p_i, we can write

$$d = p_1^{f_1} \cdots p_r^{f_r},$$

where $f_i \leq e_i$ for all i. Then $d|b$ if and only if $f_i \leq e'_i$, $1 \leq i \leq r$. Conversely, if $f_1, \cdots, f_r \in \mathbb{Z}_{\geq 0}$, then d as above divides a and b if and only if $f_i \leq e_i$, and $f_i \leq e'_i$. It follows that

$$(a, b) = p_1^{\min(e_1, e'_1)} p_2^{\min(e_2, e'_2)} \cdots p_r^{\min(e_r, e'_r)}. \tag{3}$$

On the other hand, $l \in \mathbb{N}$ is a common multiple of a and b, if and only if for each $p = p_i$, the prime factorization of l has a factor p^e with $e \geq e_i$, $e \geq e'_i$. It follows that

$$lcm(a, b) = p_1^{\max(e_1, e'_1)} p_2^{\max(e_2, e'_2)} \cdots p_r^{\max(e_r, e'_r)}. \tag{4}$$

Using prime factorizations, we can immediately deduce several divisibility properties. It is convenient to introduce the following notation. For p prime, $e \in \mathbb{Z}_{\geq 0}$, we say p^e *exactly divides* n, if $p^e \mid n$, and $p^{e+1} \nmid n$. In other words, p^e exactly divides n if and only if the exponent of p in the prime factorization of n is equal to e. We write this as: $p^e \| n$.

Proposition 3.4. *Suppose $a, b, n \in \mathbb{N}$.*

(i) *If $(a, b) = 1$, $a|n$, $b|n$, then $ab|n$.*
(ii) *$lcm(a, b) \cdot gcd(a, b) = a \cdot b$*
(iii) *If m is a common multiple of a and b, $lcm(a, b) \mid m$.*
(iv) *If $(n, a) = 1$, then $(n, b) = (n, ab)$.*

Proof. (*i*) Let p be a prime divisor of ab. Suppose that for $e, e', e'' \in \mathbb{Z}_{\geq 0}$, $p^e \| a$, $p^{e'} \| b$, $p^{e''} \| n$. Then we have

$$\left.\begin{array}{l} a \mid n \Longrightarrow e \leq e'' \\ b \mid n \Longrightarrow e' \leq e'' \end{array}\right\} \Longrightarrow \max(e, e') \leq e''$$

Now

$$(a, b) = 1 \Longrightarrow e = 0 \text{ or } e' = 0 \Longrightarrow \max(e, e') = e + e'.$$

It follows that $e + e' \leq e''$. As this holds for all prime divisors of ab, $ab \mid n$.

(ii) Suppose $p^e \| a$, $p^{e'} \| b$, with p a prime. The proposition follows from the fact that since $e, e' \geq 0$,

$$\min(e, e') + \max(e, e') = e + e'.$$

(iii) By (ii) if $(a, b) = 1$, $lcm(a, b) = ab$. Then (iii) follows from (i).

(iv) Assume $(n, a) = 1$, and suppose $d|n$. If $d \mid b$, clearly $d \mid ab$. Suppose that conversely $d \mid ab$, and let p be a prime factor of d, with $p^e \| d$. Since $p \mid n$, and $(a, n) = 1$, $p \nmid a$. Then $(p^e, a) = 1$, which together with $p^e \mid ab$ implies $p \mid b$. This shows the common factors of n and b are the same as the common factors of n and ab. □

4. Quadratic Diophantine Equations, Infinitude of Primes

4.1. **Diophantine Equations.** A diophantine equation is a polynomial equation in several variables, whose integer solutions are of interest. We have already seen how to solve diophantine equations of the form

$$ax + by = c,$$

where $a, b, c \in \mathbb{Z}$, and x, y are variables.

Recall the equation

$$x^2 + y^2 = z^2 \tag{5}$$

for the three sides of a right-angled triangle.

Definition 4.1. A **Pythagorean triple** is a solution (x, y, z) of $x^2 + y^2 = z^2$, where $x, y, z \in \mathbb{Z}$. It is called **primitive** if $gcd(x, y, z) = 1$.

Note that if a prime divides any two of x, y, z, then $x^2 + y^2 = z^2$ implies it also divides the third. Then for Pythagorean triples $gcd(x, y, z) = 1$ is equivalent to x, y, z being pairwise relatively prime.

Suppose for some $x, y, z \in \mathbb{Z}$, $gcd(x, y, z) = g > 1$, so that $x = x'g$, $y = y'g$, $z = z'g$, for $x', y', z' \in \mathbb{Z}$. Then (x, y, z) is a Pythagorean triple if and only if (x', y', z') is a primitive one. Therefore in solving (5) it's enough to find all the primitive solutions.

Solution 1 (Algebraic):

Suppose (x, y, z) is a primitive Pythagorean triple, and assume (wolog) that $x, y, z > 0$. Evidently exactly two among x, y, z must be odd. In fact the only possibilities are $(x, y, z) = (\text{odd}, \text{even}, \text{odd})$ and $(\text{even}, \text{odd}, \text{odd})$. It's safe to assume (wolog) x is even, and to put $x = 2x_0$. Then

$$4x_0^2 = z^2 - y^2 = (z - y)(z + y)$$

Now we have $(z - y, z + y) = (z - y, 2y)$ (by Corollary 1.9). Since $z - y$ is even, and y is odd, we have

$$(z - y, 2y) = 2(\frac{z - y}{2}, y) = 2(z - y, y) = 2(z, y) = 2.$$

Writing $z - y = 2k$, $z + y = 2k'$, we have $x^2 = kk'$, where $(k, k') = 1$. If we now consider the prime factorization of x^2, we see that if $p^{2e} \| x^2$, then either $p^{2e} \| k$ and $p \nmid k'$, or the other way around. It follows that k and k' are both squares, up to sign. First assume $k, k' > 0$, so that for $m, n \in \mathbb{Z}$,

$$z - y = 2n^2, \quad z + y = 2m^2.$$

Then

$$z = m^2 + n^2, \quad y = m^2 - n^2,$$

and

$$x^2 = (m^2 + n^2)^2 - (m^2 - n^2)^2 = 4m^2n^2 \Longrightarrow x = \pm 2mn.$$

Note that replacing m by $-m$ changes $2mn$ to $-2mn$. Since m, n vary over all integer, we can just write $x = 2mn$ and parametrize the solutions by

$$(x, y, z) = (2mn, m^2 - n^2, m^2 + n^2), \quad m, n \in \mathbb{Z} \tag{6}$$

Our assumptions that (x, y, z) is primitive and $x, y, z > 0$ imply $(m, n) = 1$ and $m > n$, but the above formulas give valid solutions for all $m, n \in \mathbb{Z}$. The case $k, k' < 0$ multiplies $z + y$ and $z - y$ by -1, with x^2 remaining the same. We can therefore write all the pythagorian triples as

$$\left\{\begin{array}{c} x = 2mn, \\ y = m^2 - n^2 \\ z = \pm(m^2 + n^2) \end{array}\right. , \quad \text{and} \quad \left\{\begin{array}{c} x = m^2 - n^2, \\ y = 2mn \\ z = \pm(m^2 + n^2) \end{array}\right. , \quad m, n \in \mathbb{Z}.$$

Here's now a different solution.

Solution 2 (Geometric):

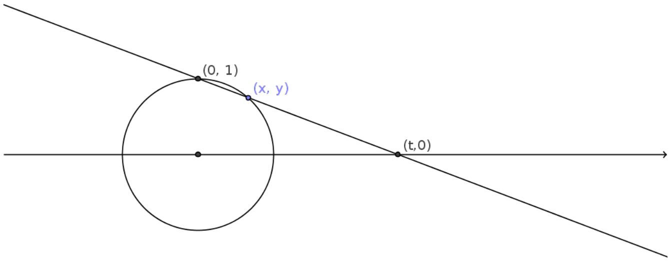
We note that in the equation $x^2 + y^2 = z^2$, we have $z \neq 0$ unless $(x, y, z) = (0, 0, 0)$. Then setting $X = x/z$ and $Y = y/z$, we can divide the equation by z^2 to obtain

$$X^2 + Y^2 = 1. \tag{7}$$

Then we have a *rational* point (X, Y) on the unit circle. Conversely, suppose $X, Y \in \mathbb{Q}$ satisfy (7). We can write $X = x/z_1$ and $Y = y/z_2$ in lowest terms, with z_1, $z_2 > 0$.

Exercise: Prove that $z_1 = z_2$ necessarily.

The we have $X = x/z$ and $Y = y/z$ for a unique primitive triple (x, y, z), with $z > 0$. In other words, the *rational* points (X, Y) of the unit circle are in bijection with such primitive Pythagorean triples (x, y, z).



Now consider the line passing through X, Y and $(0, 1)$. It intersects the X-axis at some point $(t, 0)$. Since $(t, 0)$ also determines (X, Y), there's a bijection between the unit circle minus the point $(0, 1)$, and the x-axis.

Suppose now that $t > 1$, as in the picture. Since the line segment $(0, 1) - (X, Y)$ has the same slope as $(0, 1) - (t, 0)$, we have

$$\frac{Y}{X - t} = \frac{-1}{t} \Longrightarrow Y = 1 - \frac{X}{t},$$

which plugging in (7) and using $X \neq 0$ gives

$$t = \frac{X}{1-Y}, \quad X = \frac{2t}{t^2+1}, \quad Y = \frac{t^2-1}{t^2+1}.$$

These relations show that $X, Y \in \mathbb{Q}$ if and only if $t \in \mathbb{Q}$. In other words, under the bijection $(X, Y) \leftrightarrow t$, the *rational* points (X, Y) of the unit circle correspond to *rational points* $(t, 0)$ on the x-axis.

Given $t \in \mathbb{Q}$, write $t = \frac{m}{n}$ with $(m, n) = 1$, $n > 0$. We then have

$$(2t)^2 + (t^2-1)^2 = (t^2+1)^2 \Longrightarrow (2mn)^2 + (m^2-n^2)^2 = (m^2+n^2)^2.$$

For $t > 0$, the above solutions correspond to one-half of the circle. The other half will correspond to the second formula we obtained from the algebraic method. Note that although the idea of this method was geometric, but the actual solution involved only algebra.

This geometric method works in a large number of cases of the form

$$ax^2 + by^2 = cz^2,$$

where a, b, c are non-zero integers. Using $X = x/z$ and $Y = y/z$ one rewrites as

$$aX^2 + bY^2 = c,$$

which is the equation of either a circle, ellipse, or a hyperbola. Starting with one rational solution (X_0, Y_0), one establishes a bijection between every other rational solution (X, Y) and the points $(t, 0)$ on the X-axis, with $t \in \mathbb{Q}$.

Infinitude of Primes. If we make a long list $1, 2, 3, \cdots$ of natural numbers, and mark out the prime ones, we can hardly avoid guessing there must be infinitely many primes. Indeed, primes appear much more frequently in such a list than say, square numbers. Nevertheless the infinitude of primes requires a proof. The oldest proof is also the simplest.

Theorem 4.2 (Euclid). *There are infinitely many primes.*

Proof. Let $p_1, \cdots, p_n$ be a finite list of primes, and put $A = p_1 \cdot p_2 \cdots p_n + 1$. As $A > 1$, it is divisible by *some* prime, which can't be any of $p_1, \cdots, p_n$ because dividing A by any of those gives the remainder 1. Therefore $\{p_1, \cdots, p_n\}$ can not be a complete list of primes. □

Although Euclid's proof is typically formulated as a proof by contradiction, it actually prescribes an algorithm for generating an infinite list of primes, though not quite a practical one as such. Let $p_1 = 2$ and define p_n inductively as the smallest prime factor of $p_1 \cdots p_{n-1} + 1$. This is a well-defined, infinite list of primes, beginning with

$$2,\ 3,\ 7,\ 43,\ 13,\ 53,\ 5,\ 6221671,\ 38709183810571,\ 139,\ 2801,\ 11,\ 17,\ 5471, \cdots.$$

There are several other proofs of the infinitude of primes with more or less the same underlying idea. Here's a sketch of a completely different proof, due to Euler, coming about two thousand years after Euclid.

Consider the function $\zeta(s)$, defined by the infinite series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots.$$

It's not too difficult to prove that $\zeta(1)$ is a divergent series, and by some basic calculus that $\zeta(s)$ converges for $s > 1$. For such values of s, $\zeta(s)$ can be written as a product over all primes:

$$\zeta(s) = \prod_{p \text{ prime}} (1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots).$$

Indeed, if $n \in \mathbb{N}$ has prime factorization

$$n = p_1^{\alpha_1} \cdots p_r^{\alpha_r},$$

we have

$$\frac{1}{n^s} = \frac{1}{p_1^{\alpha_1 s}} \cdot \frac{1}{p_2^{\alpha_2 s}} \cdots \frac{1}{p_r^{\alpha_r s}}.$$

As prime factorization is unique, the expression above occurs exactly once in the product

$$(1 + \frac{1}{2^s} + \frac{1}{2^{2s}} + \frac{1}{2^{3s}} \cdots)(1 + \frac{1}{3^s} + \frac{1}{3^{2s}} + \cdots)(1 + \frac{1}{5^s} + \frac{1}{5^{2s}} + \cdots) \cdots .$$

Conversely, each term of the expansion of this product is of the form $\frac{1}{n^s}$ for some $n \in \mathbb{N}$.

Then summing the geometric series

$$\frac{1}{1 - p^{-s}} = 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots ,$$

we can write

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}}. \tag{8}$$

If there only existed finitely many primes, (8) would be a product with finitely many factors, hence $\zeta(s)$ would have a finite value for all s. But that contradicts the fact that $\zeta(s) \to \infty$ as $s \to 1^+$, so there must be infinitely many primes.

This account of the proof is not quite rigorous, and neither was Euler's version. But the main idea is sound and it can be made strictly correct by some careful analysis. It's worth mentioning however, the striking fact implied by the very possibility of such a proof: that techniques from analysis can be applied to prove facts about prime numbers. This idea is the starting point of *analytic number theory*. The topic is well outside the scope of this course, so we won't delve any further into it.

5. Modular Arithmetic, Wilson's Theorem

Modular Arithmetic

Let $m \in \mathbb{N}$.

Definition 5.1. For $a, b \in \mathbb{N}$, we say a **is congruent to** b **modulo** m, if $m|a - b$. We write:

$$a \equiv b \text{ (mod } m).$$

(Common variations are $a \overset{m}{\equiv} b$, $a \equiv b$ (m), or just $a \equiv b$ if m is understood.)

If $a = qm + r$, with $0 \leq r < m$, then $a \equiv r$ (mod m). Suppose $b = q'm + s$, $0 \leq s < m$, so that also $b \equiv s$ (mod m). Then $a - b = m(q - q') + r - s$, so that

$$m|a - b \iff m|r - s.$$

But $-m < r - s < m$, so $m|r - s$ if and only if $r - s = 0$. Hence

$$a \equiv b \text{ (mod } m) \iff a \text{ and } b \text{ have the same remainder after division by } m.$$

Proposition 5.2. *Congruence modulo m is an equivalent relation. In other words for all $a, b, c \in \mathbb{Z}$:*

(1) $a \equiv a$ *(mod m)*
(2) $a \equiv b$ *(mod m)* $\iff$ $b \equiv a$ *(mod m)*
(3) $a \equiv b$ *(mod m)*, $b \equiv c$ *(mod m)* $\implies$ $a \equiv c$ *(mod m)*.

Proof. Left as exercise. □

The equivalence relation signified by $\equiv \pmod{m}$ leads to a partition of $\mathbb{Z}$ into m disjoint *equivalence classes*, represented by the remainders $0, 1, \cdots, m-1$. In other words, each $a \in \mathbb{Z}$ is congruent modulo m to exactly one element of $S = \{0, \ \cdots, \ m-1\}$. We will sometimes refer to the classes represented by these remainders as *residues*.

The following proposition says that congruence modulo m is compatible with addition and multiplication.

Proposition 5.3. *If $a \equiv a'$, $b \equiv b'$ (mod m), then*

$$a + b \equiv a' + b' \text{ and } ab \equiv a'b'.$$

Proof. If $a \equiv a' \pmod{m}$, $a - a' = mk$ for some k. Likewise $b \equiv b'$ implies $b - b' = mk'$ for some k'. Then

$$a + b - (a' + b') = m(k + k'),$$

so $a + b \equiv a' + b'$. And

$$a'b' - ab = (a - mk)(b - mk') - ab = m(mkk' - ak' - bk),$$

so $ab \equiv a'b'$. □

Some immediate applications:

Proposition 5.4. *For all $n \in \mathbb{Z}$:*

- $n^2 \equiv 0$ *or* 1 (*mod* 3).
- $n^2 \equiv 0$ *or* 1(*mod* 4). *In fact if n is odd,* $n^2 \equiv 1$(*mod* 8).
- $n^2 \equiv 0, 1,$ *or* -1 (*mod* 5)

Proof. Since either $n \equiv 0, 1,$ or $2 \pmod{3}$, the first assertion follows from $2^2 = 4 \equiv 1 (\text{mod } 3)$.

If n is even, i.e. $n = 2k$, then $4 | n^2 = 4k$, so $n^2 \equiv 0 \pmod{4}$. Otherwise, $n = 2k + 1$, hence

$$n^2 = 4k^2 + 4k + 1 = 4k(k + 1) + 1,$$

But since one of k or $k + 1$ must be even,

$$2 | k(k + 1) \Longrightarrow 8 | 4k(k + 1) \Longrightarrow n^2 \equiv 1 \pmod{8}.$$

Finally, modulo 5 the square residues are $0, 1, 4, 9, 16$, which are congruent to $0, 1, -1, -1, 1$. □

Example: The number 54432101326 can *not* be a perfect square, since $8 \mid 1000$ implies:

$$54432101326 = 54432101 \cdot 1000 + 326 \equiv 326 \equiv 6 \pmod{8},$$

and $6 \not\equiv 0, 1 (\text{mod } 8)$.

Note if a and b are both 1 (mod 4), so is ab. Thus if some $n \in \mathbb{N}$ has a prime factorization $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$, and if every $p_i \equiv \ 1(\text{mod } 4)$, then $n \equiv 1(\text{mod } 4)$. Equivalently, if $n \equiv 3(\text{mod } 4)$, there must be some prime $p \mid n$, such that $p \equiv 3 \pmod{4}$.

Proposition 5.5. *There are infinitely many primes p congruent to* 3 (*mod* 4).

Proof. We can adapt Euclid's proof of infinitude of primes (Theorem 4.2) as follows. Suppose $\{p_1, \cdots, p_r\}$ is a set of distinct primes congruent to 3 modulo 4. Let $n = 4p_1 \cdot p_2 \cdots p_r - 1$. Then $n \equiv -1 \equiv 3 \pmod{4}$, so it must have a prime factor p that's also congruent to 3. But none of $p_1, \cdots, p_r$ divide n, so $p \in \{p_1, \cdots, p_r\}$. Then no such finite list $\{p_1, \cdots, p_r\}$ of primes that are congruent to 3 (mod 4) can be complete. □

As in Theorem 4.2, although the proof as phrased is by contradiction, the underlying idea is an inductive procedure that generates an infinite number of primes congruent to 3 modulo 4. One puts $p_1 = 3$, and defines p_k, for $k > 1$, as the smallest prime divisor of $4p_1 \cdots p_{k-1} - 1$ that's congruent to 3 (mod 4). The first few are:

$$3,\ 11,\ 131,\ 17291,\ 298995971,\ 8779,\ 594359,\ 59,\ 151,\ 983, \ldots.$$

The primes congruent to 1 (mod 4) are also infinite, as we will later prove. In fact there is the following celebrated theorem, known as Dirichlet's theorem on primes in arithmetic progressions.

Theorem 5.6 (Dirichlet)**.** *For any modulus $m > 1$, and any $a \in \mathbb{N}$, with $(a, m) = 1$, there are infinitely many primes p congruent to a (mod m).*

The proof of this theorem is one of the crowning achievements of analytic number theory. The underlying idea is to modify Euler's analytic proof of the infinitude of primes. Of course to carry this out other new ideas are also required, for instance the zeta function $\zeta(s)$ has to be replaced with a suitable variant. All this is firmly beyond the scope of an introductory number theory course, but we can still appreciate the result.

Consider the equation

$$2 \cdot 4 \equiv 1 \pmod 7.$$

It says that modulo 7, the number 4 is multiplicative inverse of 2, and plays the role of $\frac{1}{2}$ in rational numbers.

Definition 5.7. For $m > 1$, we say a is **invertible**, or **has an inverse modulo** m, if there exists $b \in \mathbb{N}$ such that $ab \equiv 1(\text{mod } m)$. The (residue class) of the number b is then **the inverse** of a mod m. It is sometimes denoted a^{-1}, if there's no risk of confusion with $\frac{1}{a}$.

With a and b as in the definition, we have to justify saying b is *the* inverse of a, as opposed to *an* inverse.

Proposition 5.8. *For $m > 1$, if $ab \equiv ac \equiv 1$ (mod m), then $b \equiv c$.*

Proof.

$$b \equiv b \cdot 1 \equiv b(ac) \equiv (ab)c \equiv 1 \cdot c \equiv c.$$

□

Proposition 5.9. *a is invertible modulo m if and only if $(a, m) = 1$.*

Proof. We have

$$\begin{aligned} a \text{ is invertible modulo } m &\iff \exists b, ab \equiv 1 \pmod m \\ &\iff \exists b \text{ such that } m \mid ab - 1 \\ &\iff \exists b, k \text{ such that } mk = ab - 1 \\ &\iff ax + by = 1 \text{ can be solved with } x, y \in \mathbb{Z} \\ &\iff (a, b) = 1 \quad \text{(by Corollary 1.12)} \end{aligned}$$

□

Corollary 5.10. *For a prime p, a is invertible modulo p if and only if $p \nmid a$.*

It may happen that a number a is its own inverse. For example modulo 8, all the invertible numbers, i.e. $1, 3, 5$, and 7, are their own inverses. Clearly 1 and -1 are their own inverses modulo any m. Modulo a prime however, these are the only ones.

Lemma 5.11. *For a prime p, $a^2 \equiv 1$ (mod p) if and only if $p \equiv \pm 1$ (mod p).*

Proof.

$$a^2 \equiv 1 \pmod p \iff p \mid a^2 - 1 = (a+1)(a-1) \iff p \mid a-1 \text{ or } p \mid a+1 \iff a \equiv 1 \text{ or } a \equiv -1 \pmod p$$

□

We can now prove one of the classic theorems of modular arithmetic.

Theorem 5.12 (Wilson's Theorem). *For a prime p, we have*

$$(p-1)! \equiv -1 \ (mod\ p).$$

Conversely, any integer p with this property is prime.

Proof. First note that if $(m-1)! \equiv -1 \pmod{m}$, then $(m-1)!$ is invertible modulo m. Then $(m-1)!$ must be coprime to m by Proposition 5.9. In particular m is not divisible by any of $1, \cdots, m-1$, so it must be prime.

Now let p be a prime. If $p = 2$, then $(p-1)! \equiv -1$ is trivial, so assume p is odd. We know that every element of $S = \{1, \cdots, p-1\}$ is invertible by Corollary 5.10. By Lemma 5.11, every element $a \in S$ besides ± 1 has an inverse $b \in S$ such that $a \not\equiv b$. Then $S - \{-1, 1\}$ can be partitioned into two disjoint subsets S', S'', of equal size, such that the elements in S'' are exactly the inverses of those in S' modulo p. Then

$$(p-1)! = 1 \cdot (-1) \cdot \prod_{a \in S'} a \cdot \prod_{b \in S''} b \equiv (-1) \cdot \prod_{a \in S'} a \cdot \prod_{a \in S'} a^{-1} \equiv -1 \pmod{p}.$$

□

We will eventually prove a generalization of this theorem due to Gauss.

6. Fermat's Little Theorem, Euler's Theorem, and the Chinese Remainder Theorem

The following theorem, though easy to prove, is one of the central results of elementary number theory.

Proposition 6.1 (Fermat's Little Theorem). *Let p be a prime, and suppose $a \in \mathbb{N}$, $p \nmid a$. Then*

$$a^{p-1} \equiv 1 \ (mod\ p).$$

Proof. Consider the set

$$S = \{1, 2, \cdots, p-1\}$$

which consists of all non-zero residue classes modulo p. Fix $a \in \mathbb{N}$, and define a function

$$f : S \to S, \quad f(x) = ax \pmod{p}.$$

Since a is invertible residue modulo p, the function f is *injective*:

$$f(x) = f(y) \Longrightarrow ax \equiv ay \pmod{p} \Longrightarrow a^{-1}ax \equiv a^{-1}ay \pmod{p} \Longrightarrow x \equiv y \pmod{p}.$$

Since S is a finite set, injectivity of f implies it is in fact *bijective*. In other words, that the set $f(S) = \{f(b) : b \in S\}$ is the same as S. Now we compute the product of the elements in S two different ways, modulo p:

$$\prod_{s \in S} b \equiv \prod_{s' \in f(S)} s' \pmod{p}.$$

But

$$\prod_{s' \in f(S)} s' = \prod_{s \in S} f(s) = \prod_{s \in S} as = a^{p-1} \prod_{s \in S} b,$$

so we obtain

$$a^{p-1} \prod_{s \in S} s \pmod{p} \equiv \prod_{s \in S} s.$$

Now every number $s \in S$ is invertible modulo p, hence so is the product $P = \prod_{s \in S} s$. Multiplying the above by $P^{-1} \pmod{p}$, we obtain

$$a^{p-1} \equiv 1 \pmod{p}.$$

□

Example 6.2. Recalling Example 6 Let us find the last digit of 7^{2017}. First applying Fermat's theorem to $p = 5$, we have $7^4 \equiv 1 \pmod 5$. Writing $2017 = 4 \cdot 504 + 1$ we have

$$7^{2017} = (7^4)^{504} \cdot 7 \equiv 1^{504} \cdot 2 \equiv 2 \pmod 5.$$

Therefore, $7^{2017} \equiv 2$ or 7 modulo 10. Now $7^{2017} - 2$ is odd, so not divisible by 10, so $7^{2017} \equiv 7 \pmod{10}$. In other words, the final digit of 7^{2017} is 7.

Note in the example that 7 is invertible modulo 10, so that

$$7^{2017} \equiv 7 \pmod{10} \Longrightarrow 7^{2016} \equiv 1 \pmod{10}.$$

We might see that $9 \mid 2016$ and wonder if $7^9 \equiv 1 \pmod{10}$ as in Fermat's theorem, even though 10 is not prime. However,

$$7^2 \equiv 49 \equiv 9 \Longrightarrow 7^8 \equiv 81 \equiv 1 \Longrightarrow 7^9 \equiv 7 \pmod{10}$$

So one can't hope to replace p by any number m in the statement of Fermat's theorem. We can take look at the proof and see where it breaks down for non-primes.

Let $m \in \mathbb{N}$, and put

$$S = \{1, \cdots, m-1\}.$$

Then if $a \in S$ is arbitrary, the function

$$f : S \to S, \quad f(a) = ax$$

is *not* injective, unless a is invertible modulo m, a condition that is automatic for a prime modulus. Then let us assume $(a, m) = 1$ so this is the case. Then f will again be bijective, and taking the product of $s \in S$ two different ways, we obtain

$$(m-1)! \equiv a^{m-1}(m-1)! \pmod m.$$

But now we can not invert $(m-1)!$ modulo m. In fact if m is not prime, $(m-1)!$ will never be invertible, as we saw in the proof of Wilson's theorem.

The solution is to restrict to the subset

$$T = \{s \in S : (s, m) = 1\}, \tag{9}$$

of invertible elements in S.

Definition 6.3. The number of distinct invertible residue classes modulo m, is denoted $\varphi(m)$. This is called Euler's φ-function, or Euler's **totient** function.

Notice that $\varphi(m)$ is simply the size of the set T in (9). We can now state and the right generalization of Fermat's theorem.

Theorem 6.4 (Euler). *Let $m > 1$ and suppose $(a, m) \equiv 1$. Then*

$$a^{\varphi(m)} \equiv 1 \ (\mathit{mod}\ m).$$

Proof. Proceeding as in the proof of Fermat's theorem, we consider the function $f : T \to T, f(t) = at$, where T is as in (9), the set of all invertible residue classes mod m. Since a is invertible, f is bijective, and we have:

$$\prod_{t \in T} t \equiv \prod_{t \in T} f(t) \equiv \prod_{t \in T} at \equiv a^{\varphi(m)} \prod_{t \in T} t \pmod m.$$

Since each $t \in T$ is invertible mod m, so is $P = \prod_{t \in T} t$. Multiplying both sides of the above by P^{-1}, we obtain the theorem. □

Example 6.5. We revisit Example 6.2. Checking that between $1, 2, \cdots, 9$ only $\{1, 3, 7, 9\}$ are relatively prime to 10, we have $\varphi(10) = 4$. Indeed, modulo 10,

$$7^4 \equiv 49^2 \equiv (-1)^2 = 1.$$

So

$$7^{2016} \equiv 1^{504} \equiv 1 \pmod{10} \implies 7^{2017} \equiv 7 \pmod{10}.$$

Then the correct explanation that $7^{2016} \equiv 1 \pmod{10}$ is not $10 - 1 = 9$ divising 2016, but rather $\varphi(10) = 4$ dividing it.

Example 6.6. Suppose we want to find the remainder of 3^{30} after division by 35.

Method 1: Using Euler's Theorem.

We have

$$\varphi(35) = \#\{1, 2, \cdots, 34\} - \#\{5, 10, 15, 20, 25, 30, 7, 14, 21, 28\} = 34 - 10 = 24.$$

Here we have listed every natural number smaller than 35 that shares a factor with it, but later we will see that there are better ways. Now

$$3^{\varphi(35)} \equiv 3^{24} \equiv 1 \pmod{35}.$$

$$\implies 3^{30} \equiv 3^6 = 3^4 \cdot 3^2 \equiv 11 \cdot 9 = 99 \equiv 29 \pmod{35}.$$

Method 2: Using Fermat's Theorem.

Note that $35 = 7 \cdot 5$. Then

$$3^4 \equiv 1 \pmod 5 \implies 3^{30} = 3^{28} \cdot 3^2 \equiv 4 \pmod 5,$$

$$3^6 \equiv \ 1 \pmod 7 \implies 3^{30} \equiv 1 \pmod 7.$$

If a is the remainder of 3^{30} after division 35,

$$a \equiv 4 \pmod 5 \implies a \in \{4, 9, 14, 19, 24, \underline{29}, 34\}, \text{ and}$$

$$a \equiv 1 \pmod 7 \implies a \in \{1, 8, 15, 22, \underline{29}.\}$$

So a must be the number in common between the two sets, which is 29. Therefore $3^{30} \equiv 29 \pmod{35}$.

In Example 6.6 we saw that there was only one residue class modulo 35, that is $\equiv 4 \pmod 5$ and $\equiv 1 \pmod 7$. This is not an accident.

Proposition 6.7. *Suppose $(m, n) = 1$, with $m, n > 0$, and $a, b \in \mathbb{Z}$. Then the system of congruences*

$$x \equiv a \ (mod\ n)$$

$$x \equiv b \ (mod\ m)$$

has a <u>*unique*</u> *solution modulo mn.*

Proof. First we show uniqueness of the solution, assuming it exists. Suppose x and y were two solutions of the congruence system. Then $x \equiv y \pmod m$ implies $m \mid x - y$, and $x \equiv y \pmod n$ implies $n \mid x - y$. Then $(m, n) = 1$ implies $mn \mid x - y$, which is to say $x \equiv y \pmod{mn}$.

Now we show the *existence* of a solution. Since $(m, n) = 1$, there are integers $\alpha, \beta \in \mathbb{Z}$ such that $\alpha m + \beta n = 1$ (Corollary 1.12). Note that,

$$\alpha m \equiv \begin{cases} 1 & \pmod n \\ 0 & \pmod m \end{cases} \qquad \beta n \equiv \begin{cases} 0 & \pmod n \\ 1 & \pmod m \end{cases}$$

Then

$$a(\alpha m) + b(\beta n) \equiv \begin{cases} a & \pmod n \\ b & \pmod m), \end{cases}$$

so $x = a(\alpha m) + b(\beta n)$ is the unique solution modulo mn. □

The proposition is a special case of a more general theorem.

Definition 6.8. Integers $m_1, \cdots, m_r \in \mathbb{Z}$ are said to be *pairwise relatively prime*, or *pairwise coprime*, if $(m_i, m_j) = 1$ for all $i \neq j$.

Theorem 6.9 (Chinese Remainder Theorem)**.** *Suppose $m_1, \cdots, m_r$ are pairwise relatively prime, and that $a_1, \cdots a_r \in \mathbb{Z}$. Then the system of congruences*

$$\begin{gathered} x \equiv a_1 \ (mod\ m_1) \\ x \equiv a_2 \ (mod\ m_2) \\ \vdots \\ x \equiv a_r \ (mod\ m_r) \end{gathered}$$

has a unique integer solution for x modulo $m = m_1 m_2 \cdots m_r$.

Proof. We proceed by induction. If $r = 1$, the statement is trivial. Now assume the theorem holds for $r = k$. Given $a_1, \cdots, a_{k+1}$ and $m_1, \cdots, m_{k+1}$ as in the statement, let $m' = m_1, \cdots, m_k$. Then by the assumption there exists a unique a' modulo m' such that

$$x \cong a_i \ (\text{mod } m_i),\ 1 \leq i \leq k \iff x \cong\ a' \ (\text{mod } m).$$

Therefore the system is equivalent to

$$x \equiv a' \ (\text{mod } m'), \quad x \equiv a_{k+1} \ (\text{mod } m_{k+1}).$$

where $(m', m_{k+1}) = 1$ (by Proposition 3.4(iv), for instance). Now the theorem follows the case $r = 2$, which is Proposition 6.7. □

Suppose $m, n > 1$ and $(m, n) = 1$. The set

$$R_m = \{1, 2, \cdots, m\}, \tag{10}$$

is a complete set of residues modulo m. Let R_n be defined the same way, and consider the function

$$\psi : \mathbb{Z} \to R_m \times R_n, \quad x \mapsto (x \ (\text{mod } m), x \ (\text{mod } n)).$$

By the Chinese Remainder Theorem, given $a \in \mathbb{Z}$, knowledge of $\psi(a)$ determines a modulo mn (and vice versa), so that

$$\psi(x) = \psi(y) \iff x \equiv y \ (\text{mod } mn).$$

Example 6.10. Let $m = 9$, $n = 14$, so that $mn = 126$. The numbers $a = 1, \cdots, 126$ fit into a 9×14 table as follows.

28	56	84	112	14	42	70	98	126
55	83	111	13	41	69	97	125	27
82	110	12	40	68	96	124	26	54
109	11	39	67	95	123	25	53	81
10	38	66	94	122	24	52	80	108
37	65	93	121	23	51	79	107	9
64	92	120	22	50	78	106	8	36
91	119	21	49	77	105	7	35	63
118	20	48	76	104	6	34	62	90
19	47	75	103	5	33	61	89	117
46	74	102	4	32	60	88	116	18
73	101	3	31	59	87	115	17	45
100	2	30	58	86	114	16	44	72
1	29	57	85	113	15	43	71	99

The columns are the distinct residue classes modulo 9 and the rows residue classes modulo 14. This table represents the solution to the system of congruences

$$x \equiv a \pmod{9}, \quad y \equiv b \pmod{14},$$

where a is the column number from the left, and b is the row number from the bottom. Note how we can construct this by writing $1, 2, \cdots, 126$ into the table starting in the bottom left and moving up and right each step, wrapping around the edges whenever they are reached. As the Chinese Remainder Theorem predicts, since 14 and 9 are relatively prime, all 126 numbers fit exactly into the grid without overlap.

Suppose we want to compute $\varphi(mn)$. We might construct a table such as the one above, and cross out all numbers $a \in \{1, \cdots, mn\}$ that share a factor in common with mn. Such an a would have a factor common either with m or with n. If $(a, m) > 1$, then every number in the same column as a would be crossed out along with a, since $(a, m) = (a \pm m, m)$. Similarly, if $(a, n) = 1$ every other number in the same row would also be crossed. Then in fact we would be crossing out entire rows and columns.

For instance if $a = 7$ in the example of $m = 9$ and $n = 14$, the entire row containing 7 would be crossed out. After crossing out all such rows and columns the result will look like:

28	56	84	112	14	42	70	98	126
55	83	111	13	41	69	97	125	27
82	110	12	40	68	96	124	26	54
109	11	39	67	95	123	25	53	81
10	38	66	94	122	24	52	80	108
37	65	93	121	23	51	79	107	9
64	92	120	22	50	78	106	8	36
91	119	21	49	77	105	7	35	63
118	20	48	76	104	6	34	62	90
19	47	75	103	5	33	61	89	117
46	74	102	4	32	60	88	116	18
73	101	3	31	59	87	115	17	45
100	2	30	58	86	114	16	44	72
1	29	57	85	113	15	43	71	99

If we now delete the rows and columns that contain numbers not coprime to 126, the result will be a smaller rectangle:

55	83	13	41	97	125
109	11	67	95	25	53
37	65	121	23	79	107
19	47	103	5	61	89
73	101	31	59	115	17
1	29	85	113	43	71

In general there will be $\varphi(m)$ columns left, representing residues coprime to m, and similarly $\varphi(n)$ rows, for each invertible residue class mod n. On the other hand the total number of remaining entries is $\varphi(mn)$ by construction. Therefore:

Proposition 6.11. *If* $(m, n) = 1$*,*

$$\varphi(mn) = \varphi(m)\varphi(n).$$

Proof. Here is a more formal presentation of the same proof. Suppose $m, n > 1$, $(m, n) = 1$, and R_m, R_n, and R_{mn} are complete sets of residue classes as in (10). By the Chinese Remainder Theorem, the function $\psi : R_{mn} \to R_m \times R_n, x \mapsto (x \pmod{m}, y \pmod{n})$ is bijective. Now consider the subset $S_m \subset R_m$ of invertible residue classes, i.e.

$$S_m = \{a \in R_m : (a, m) = 1\}.$$

For $c \in \mathbb{Z}$, we have $(c, mn) = (c, m) \cdot (c, n)$ since $(m, n) = 1$. Then $(c, mn) = 1 \iff (c, m) = 1$ and $(c, n) = 1$. In other words, $c \in S_{mn}$ if and only if $c \pmod{m} \in S_m$ and $c \pmod{n} \in S_n$. Then $\psi : R_{mn} \to R_m \times R_n$ restricts to a bijection $S_{mn} \to S_m \times S_n$, and

$$\varphi(mn) = |S_{mn}| = |S_m| \cdot |S_n| = \varphi(m)\varphi(n).$$

□

Example 6.12. In Example 6.6 we computed $\varphi(35) = 24$ by listing all natural numbers ≤ 35 which have a factor in common with it. Using Proposition 6.11, we can simply write

$$\phi(35) = \phi(5)\phi(7) = 4 \cdot 6 = 24.$$

Proposition 6.13. (i) *For a prime p and $\alpha \geq 1$, $\phi(p^\alpha) = p^{\alpha-1}(p-1)$.*

(ii) *If $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$ is the prime factorization of n,*

$$\varphi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

Proof. We have $(a, p^\alpha) = 1$ if and only if $p \nmid a$. For $a \in \{1, \cdots, p^\alpha\}$, we have $p \mid a$ if and only if $a \in \{p, 2p, \cdots, p^{\alpha-1} \cdot p\}$. It follows that $\varphi(p^\alpha) = p^\alpha - p^{\alpha-1} = p^{\alpha-1}(p-1)$. This proves (i).

For (ii), using Proposition 6.11 and (i), we have

$$\varphi(n) = \prod_{p|n} \varphi(p^\alpha) = \prod_{p|n} p^\alpha - p^{\alpha-1} = \prod_{p|n} p^\alpha \left(1 - \frac{1}{p}\right) = n \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

□

7. Polynomial Equations modulo n, Hensel's Lemma

Suppose that for $f(x) \in \mathbb{Z}[x]$, and $n > 1$, we want to solve

$$f(x) \equiv 0 \pmod{n}.$$

Let

$$m = p_1^{\alpha_1} \cdots p_r^{\alpha}.$$

denote the prime factorization of m. Then for $a \in \mathbb{Z}$

$$f(a) \equiv 0 \pmod{m} \implies f(a) \equiv 0 \pmod{p_i^{\alpha_i}}.$$

Conversely, suppose that $a_1, \cdots a_r \in \mathbb{Z}$ satisfy

$$f(a_i) \equiv 0 \pmod{p_i^{\alpha_i}}.$$

By the Chinese Remainder Theorem there exists a unique residue a modulo m, such that

$$a \equiv a_i \pmod{{p_i}^{\alpha_i}}.$$

In particular

$$f(a) \equiv f(a_i) \equiv 0 \pmod{p_i^{\alpha_i}},$$

which implies

$$f(a) \equiv 0 \pmod{m}.$$

This shows that there's a bijection between sets of residue classes

$$\{a : f(a) \equiv 0 \pmod{m}\} \leftrightarrow \{(a_1, \cdots, a_r) : f(a_i) \equiv 0 \pmod{p_i^{\alpha_i}}\},$$

where a is considered modulo m, and a_i considered modulo $p_i^{\alpha_i}$. Therefore solving the equation

$$f(a) \equiv 0 \ (\mathrm{mod}\ m),$$

can be reduced to the case $m = p^\alpha$, using the Chinese Remainder Theorem.

Now suppose $m = p^\alpha$, so that we are interested in solving

$$f(a) \equiv 0 \ (\mathrm{mod}\ p^\alpha).$$

A naive method of solution is to try all residues modulo p^α. As there are finitely many in total, this works in principle, but is far from ideal, and impractical as soon as α is not small.

However, one may observe that a solution of $f(x) \equiv 0 \ (\mathrm{mod}\ p^\alpha)$ is necessarily also a solution of

$$f(x) \equiv 0 \ (\mathrm{mod}\ p),$$

which is easier to solve, even by brute search. Suppose we know that $S_1 \subset \{0, 1, \cdots, p-1\}$ is the set of solutions of this equation. Then if $a \in \{0, \cdots, p^\alpha - 1\}$ satisfies $f(a) \equiv 0 \ (\mathrm{mod}\ p^\alpha)$, the class of $a \ (\mathrm{mod}\ p)$ must belong to S_1.

More generally, let S_i denote the set of mod p^i residues $a \in \{0, 1 \cdots, p^i - 1\}$. Then if $i > j$, we have a map

$$S_i \to S_j, \quad a \ (\mathrm{mod}\ p^i) \mapsto a \ (\mathrm{mod}\ p^j).$$

If $a \in S_i$ maps to $b \in S_j$, we say a is a *lift* of b. Considering now the chain of maps

$$S_\alpha \to S_{\alpha-1} \to \cdots \to S_1.$$

If $f(x) \equiv 0 \ (\mathrm{mod}\ p^\alpha)$ does have a solution $a \in S_\alpha$, its image $a_i \in S_i$ is a solution modulo p^i, for all $i \leq \alpha$. We may then attempt to solve $f(x) \equiv 0 \ (\mathrm{mod}\ p^\alpha)$ as follows:

1. Find a solution $a_1 \in S_1$ for $f(a_1) \equiv 0 \ (\mathrm{mod}\ p)$. Then setting $i = 1$,
2. Check if $i = \alpha$. If so, a_i is a solution of $f(x) \equiv 0 \ (\mathrm{mod}\ p^\alpha)$ as desired. Otherwise,
3. Find a lift $a_{i+1} \in S_{i+1}$ of $a_i \in S_i$. If this is possible, increment i and repeat (2).

It is *a priori* unclear if this procedure works, since it may not be possible to lift $a_i \in S_i$ to $a_{i+1} \in S_{i+1}$. In other words, the map $S_{i+1} \to S_i$ may not be surjective.

In fact it turns out that often (but not always) this lifting step is not only possible, but the lift is *unique*. In such cases the class of $a_i \ (\mathrm{mod}\ p^i)$ and the class of $a_{i+1} \ (\mathrm{mod}\ p^{i+1})$ uniquely determine each other.

Theorem 7.1 (Hensel's Lemma)**.** *Suppose* $f(X) \in \mathbb{Z}[X]$, $f(a) \equiv \ (\mathit{mod}\ p^\alpha)$, *and* $f'(a) \not\equiv 0 \ (\mathit{mod}\ p)$, *where* $\alpha \geq 1$ *and* p *is prime. Then*

$$f(a + tp^\alpha) \equiv 0 \ (\mathit{mod}\ p^{\alpha+1})$$

for a unique residue t *modulo* p, *given by*

$$t \equiv -\frac{f(a)}{p^\alpha f'(a)} \ (\mathit{mod}\ p).$$

The theorem says that if $a_i \in S_i$ is a solution of $f(a_i) \equiv 0 \ (\mathrm{mod}\ p^i)$, then as long as $f'(a_i) \not\equiv 0 \ (\mathrm{mod}\ p)$, there exists some $a_{i+1} \in S_{i+1}$ lifting a_i. Note that $f'(a_{i+1}) \equiv f'(a_i) \ (\mathrm{mod}\ p)$, so if the criterion applies to $a_i \ (\mathrm{mod}\ p^i)$, it automatically applies to its lift $a_{i+1} \ (\mathrm{mod}\ p^{i+1})$, which can then be further lifted to $a_{i+2} \ (\mathrm{mod}\ p^{i+2})$, and so on.

Example 7.2. Suppose we want to solve $X^2 \equiv -1$ modulo 5^α, for some α. Letting $f(X) = X^2 + 1$, the equation $f(a) \equiv 0 \ (\mathrm{mod}\ 5)$ has two solutions modulo 5, namely $a = 2, \ 3$. Now $f'(X) = 2X$, so $f'(c) \not\equiv 0 \ (\mathrm{mod}\ p^\alpha)$ if $5 \nmid c$. We have $f'(2)^{-1} \equiv -1 \ (\mathrm{mod}\ 5)$ and $\frac{f(2)}{p} = 1$, so by Hensel's Lemma setting

$$t \equiv -\frac{f(2)}{p} \cdot f'(2)^{-1} \equiv 1(\mathrm{mod}\ 5)$$

then $a + t \cdot 5 = 2 + 1 \cdot 5 = 7$ is the unique residue class modulo 25 such that

$$7^2 \equiv -1 \pmod{25}, \quad 7 \equiv 2 \pmod 5.$$

Now $f(7)/5^2 = 2$ and $f'(2)^{-1} \equiv -1 \pmod 5$, so $7 + 2 \cdot 5^2 = 57$ is the unique lift of $7 \pmod{25}$ to a solution of $x^2 \equiv -1 \pmod{125}$. One continues in this way until reaching the unique solution of $x^2 \equiv -1 \pmod{5^\alpha}$ that lifts $2 \pmod 5$. The sequence thus obtained starts with

$$2,\ 7,\ 57,\ 182,\ 2057,\ 14557,\ 45807,\ \cdots$$

Likewise, another sequence gives the unique lifts of the solution $a = 3$, beginning with

$$3,\ 18,\ 68,\ 443,\ 1068,\ 1068,\ 32318,\ \cdots.$$

Proof of Theorem 7.1. If there exists $b \pmod{p^{\alpha+1}}$ such that $b \equiv a \pmod{p^\alpha}$, then b necessarily has the form $b = a + tp^\alpha$ for some integer t, which is well-determined modulo p. We wish to show a unique such t exists under the assumptions in the theorem. Note that for any exponent m, we have

$$b^m = (a + tp^\alpha)^m = a^m + \binom{m}{1} a^{m-1} t p^\alpha + \binom{m}{2} a^{m-2} t^2 p^{2\alpha} + \cdots \binom{m}{m} t^m p^{m\alpha}.$$

Note that all but the first two terms are zero modulo $p^{\alpha+1}$:

$$b^m \equiv a^m + m a^{m-1} t p^\alpha \pmod{p^{\alpha+1}}.$$

Now if

$$f(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots c_1 x + c_0,$$

the key computation is

$$f(b) = \sum_{k=0}^{n} c_k b^k \equiv c_0 + \sum_{k=1}^{n} c_k (a^k + k a^{k-1} t p^\alpha) = \sum_{k=0}^{n} c_k a^k + t p^\alpha \sum_{k=1}^{n} k a^{k-1} = f(a) + t p^\alpha f'(a) \pmod{p^{\alpha+1}}.$$

Therefore

$$f(b) \equiv 0 \pmod{p^{\alpha+1}} \iff f(a) + t p^\alpha f'(a) \equiv 0 \pmod{p^{\alpha+1}}.$$

As $p^\alpha \mid f(a)$, the above is the same as

$$\frac{f(a)}{p^\alpha} + t f'(a) \equiv 0 \pmod p,$$

and since $f'(a)$ is invertible modulo p, equivalently

$$t \equiv -\frac{f(a)}{p^\alpha f'(a)} \pmod p.$$

□

Corollary 7.3. *Suppose $f(X) \in \mathbb{Z}[X]$ and $f(a) \equiv 0 \pmod p$ for a prime p. If $f'(a) \not\equiv 0 \pmod p$, then there is a unique sequence of residues $a_k \pmod{p^k}$, $k \geq 1$, starting with $a_1 = a$, such that $f(a_k) \equiv 0 \pmod{p^k}$, and $a_{k+1} \equiv a_k \pmod{p^k}$. They are determined recursively by*

$$a_{k+1} \equiv a_k - f(a_k) \cdot s \pmod{p^{k+1}},$$

where $s \equiv f'(a)^{-1} \pmod p$.

Proof. This is simply a restatement of Hensel's Lemma for the case $\alpha = 1$, combined with the observations that $s = f'(a)^{-1} \pmod p$ only needs to be computed once. If t is as in the theorem,

$$a + tp^\alpha \equiv a - \frac{f(a)}{p^\alpha f'(a)} p^\alpha = a - \frac{f(a)}{f'(a)} \pmod p.$$

□

Example 7.4. Consider $f(X) = X^2 + X + 1$, with $p = 13$. We have $f(3) \equiv 0 \pmod{13}$, $f'(3) = 7 \not\equiv 0 \pmod{13}$, and $f'(3)^{-1} = 7^{-1} \equiv 2 \pmod{13}$. There is then a unique sequence of lifts of $3 \pmod{13}$ to $a_k \pmod{13^k}$, $k = 1, 2, \cdots$, with $f(a_k) \equiv 0 \pmod{13^k}$, starting with

$$3,\ 146,\ 1160,\ 20933,\ 220860,\ 963446,\ 20270682,\ 271264750, \cdots .$$

Hensel's Lemma only applies to roots a of $f(x)$, such that $f'(a) \not\equiv 0 \pmod p$. To be thorough, we should investigate what happens in the other cases.

Definition 7.5. A root a of $f(a) \equiv 0 \pmod p$ is called *singular* if $f'(a) \equiv 0 \pmod p$, and otherwise is called *non-singular*.

Proposition 7.6. *Let $f(X) \in \mathbb{Z}[X]$, $a \in \mathbb{Z}$ and p be a prime, $\alpha \geq 1$. Suppose a is a singular root of $f(X) \equiv 0$ $(mod\ p^\alpha)$, i.e. $f'(a) \equiv 0$ $(mod\ p)$. Then*

$$a\ (mod\ p^\alpha) \begin{cases} \textit{has no lifts } (mod\ p^{\alpha+1}) & \textit{if } p^{\alpha+1} \nmid f(a) \\ \\ \textit{has } p \textit{ lifts } b = r + t \cdot p^\alpha,\ t = 0, \cdots, p-1 & \textit{if } p^{\alpha+1} \mid f(a) \end{cases}$$

Proof. For $b = a + tp^\alpha$ we have

$$f(b) = f(a + tp^\alpha) \equiv f(a) + tp^\alpha f'(a) \pmod{p^{\alpha+1}} \equiv f(a) \pmod{p^{\alpha+1}}.$$

Thus $f(b) \equiv 0 \pmod{p^{\alpha+1}}$ if and only if $p^{\alpha+1} \mid f(a)$, and in that case any t works. □

Example 7.7. For $f(X) = X^2 + X + 1$, we have $f(1) \equiv 0 \pmod 3$, and $f'(1) = 2 \cdot 1 + 1 \equiv 0$, so 1 is a singular root. It can not be lifted modulo 9, since $3^2 \nmid f(1) = 3$.

On the other hand $f(X) = X^2 + 2X + 1$ has a root $X = 10 \pmod{11}$, which is also singular as $f'(10) = 2 \cdot 10 + 2 \equiv 0 \pmod{11}$. Since $11^2 \mid f(10) = 121$, there are 11 numbers

$$10,\ 21,\ 32,\ 43,\ 54,\ 65,\ 76,\ 87,\ 98,\ 109,\ 120,$$

all satisfying $a^2 + 2a + 1 \equiv 0 \pmod{11^2}$, and $a \equiv 10 \pmod{11}$.

8. Polynomial modulo p

Let $m > 1$, and $f[x] \in \mathbb{Z}[x]$. Previously, we saw how solving

$$f(x) \equiv 0 \pmod m \tag{11}$$

for integers x modulo m, may be reduced to the system of equations

$$f(x) \equiv c \pmod{p_i^{e_i}}, \quad i = 1, \cdots, r$$

where $m = p_1^{e_1} \cdots p_r^{e_r}$ is the factorization into primes. Thus if we know how to solve (11) for $m = p^e$, we know how to solve it for general m.

Next we saw that by Hensel's Lemma, in most cases the solution of an equation

$$f(x) \equiv 0 \pmod{p^e}$$

may computed from a solution to

$$f(x) \equiv 0 \pmod p$$

by a recursive procedure. We are therefore interested in solving polynomial equations modulo p.

Let us first make a digression to explain the somewhat subtle difference between polynomials and functions, and the essential difference between $f(x) \equiv 0$ modulo m and modulo p.

Polynomial functions modulo m. The notion of divisibility makes sense for polynomials. For $f(X), g(X) \in \mathbb{Z}[X]$ we say f is divisible by g, and write $g(X) \mid f(X)$, if there exists $h(X) \in \mathbb{Z}[X]$ such that $f(X) = g(X)h(X)$.

As each integer $m \in \mathbb{Z}$ can be considered a constant polynomial, it makes sense to write $m|f(X)$, which amounts to saying the coefficients of $f(X)$ are divisible by m. We can also extend congruences to polynomials the same way, and write

$$f(X) \equiv g(X) \pmod m \iff m|f(X) - g(X).$$

If a is a residue class modulo m, then $f(a)$ is well-defined modulo m. In other words if S is a complete system of residue modulo m, such as $\{1, \cdots, m\}$, then each $f(X) \in \mathbb{Z}[X]$ determines a well-defined function $S \to S,\ a \mapsto f(a)$.

However, polynomials that are non-congruent modulo m may give rise to the *same function* on residue classes. Consider for instance

$$f(X) = X^6 + 3, \quad g(X) = X^2 - 2.$$

Then $f(X) \not\equiv g(X) \pmod 5$, since their coefficients are not all congruent modulo 5, and yet $f(x) \equiv g(x) \pmod 5$ for every $x \in \mathbb{Z}$, since $x^5 \equiv x \pmod 5$ by Fermat's Theorem.

Then we have to be clear about the distinction between *polynomials* and *functions*. By definition, a polynomial with real coefficients is simply a formal sum

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_i \in \mathbb{R}$. Such expressions can be added, subtracted and multiplied in predictable ways without needing to be considered as functions in any way. It so happens that such a polynomial *does* give rise to a function, via substituting the variable x with a number a. The reason we normally get away with conflating polynomials and functions is the following.

Proposition 8.1. *The map*

$$\mathbb{R}[X] \to \{\mathbb{R}\text{-valued functions on } \mathbb{R}\},$$

which to a polynomial $f(X) \in \mathbb{R}[X]$ associates the function $\mathbb{R} \to \mathbb{R},\ a \mapsto f(a)$, is injective.

Proof. Let $f(X), g(X) \in \mathbb{R}[X]$, and put $h(X) = f(X) - g(X)$. If $f(a) = g(a)$ for all $a \in \mathbb{R}$, and yet $f(X) \neq g(X)$, then $h(X)$ is a non-zero polynomial with infinitely many roots, which isn't possible. □

The proof is based on the fact that a non-zero polynomial $f(X) \in \mathbb{R}[X]$ can have at most $\deg(f)$ roots, and that $\mathbb{R}$ is an infinite field. But residues modulo m are finite in number, and in fact polynomials may have more roots modulo m than their degree. For instance if $f(X) = X^2$, then $f(a) \equiv 0 \pmod 9$ for $a = 0, 3$ and 6.

Suppose we restrict the modulus to primes p, then for

$$f(X) = X^p, \quad g(X) = X,$$

we have $f(a) \equiv g(a) \pmod p$ for all $a \in \mathbb{Z}$, but again $f(X) \neq g(X)$ as polynomials.

Proposition 8.2. *Let $f(X) \in \mathbb{Z}[X]$. Then there always exists $g(X) \in \mathbb{Z}[X]$ with* $\deg(g) < p$ *such that $f(a) \equiv\ g(a) \pmod p$ for all $a \in \mathbb{Z}$.*

Proof. For any $n > p - 1$ we can write $n = q \cdot (p-1) + r$, where $1 \leq q$ and $0 < r \leq p - 1$ (note the bounds). Then for all $a \in \mathbb{Z}$, we have

$$a^n \equiv (a^{p-1})^q \cdot a^r \equiv a^r \pmod p,$$

by Fermat's little theorem if $p \nmid a$, and trivially if $p \mid a$. Therefore modulo p the polynomial X^n takes the same values as X^r.

If $f(X) = \sum_{i=1}^{s} a_i X^{n_i} \in \mathbb{Z}[X]$, let $m_i = n_i$ if $n_i \leq p-1$ and $m_i = r_i$ if $n_i > p-1$, where r_i is obtained through dividing n_i by $p-1$ the same way as above. Then $g(X) = \sum_{i=1}^{s} a_i X^{m_i}$ satisfies the requirements of the proposition. □

Example 8.3. If $f(X) = X^{17} + X^8 - 9X^6 + X + 4$ and $p = 5$, then

$$17 = 4 \cdot 4 + 1, \quad 8 = 4 \cdot 1 + 4, \quad 6 = 4 \cdot 1 + 2.$$

So for $g(X) = X^4 - 9X^2 + 2X + 4$, we have $f(a) \equiv g(a) \pmod{p}$ for all $a \in \mathbb{Z}$.

Recall that if $f(X) \in \mathbb{R}[X]$ has degree $d > 0$ and $r \in \mathbb{R}$ satisfies $f(r) = 0$, then $f(X) = (X-r)g(X)$ for some $g(X)$ of degree $d-1$. For prime moduli there is an analogue of this fact.

Proposition 8.4. *Suppose $f[X] \in \mathbb{Z}[X]$ has degree $d > 0$ and $f(r) \equiv 0$ (mod p) for p prime, $a \in \mathbb{Z}$. Then there exists $g(X) \in \mathbb{Z}[X]$ with $\deg(g) = d-1$ such that $f(X) \equiv g(X)(X-r)$ (mod p).*

Proof. Write $f(X) = \sum_{k=0}^{d} a_k X^k$. Then

$$f(X) - f(r) = \sum_{k=0}^{d} a_k X^k - \sum_{k=0}^{d} a_k r^k = a_0 - a_0 + \sum_{k=1}^{d} a_k (X^k - r^k) = \sum_{k=1}^{d} a_k (X^k - r^k).$$

Recall that in general for $k > 0$, we have the algebraic identity

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

Let $g_k(X) = X^{k-1} + X^{k-2}r + \cdots + Xr^{k-2} + r^{k-1}$, and $g(X) = \sum_{k=1}^{d} a_k g_k(X)$. Then $\deg(g) = \deg(g_d) = d-1$, and

$$f(X) - f(r) = \sum_{k=1}^{d} a_i (X-r) g_k(X) = (X-r) \sum_{k=1}^{d} a_i g_k(X) = (X-r)g(X).$$

Since $f(r) \equiv 0 (\mathrm{mod}\ p)$, we have $f(X) \equiv (X-r)g(X) \pmod{p}$. □

Remark 8.5. Note that in the proposition we have a congruence identity between *polynomials* $f(X)$ and $g(X)(X-r)$ modulo p. This says the coefficients of $f(X)$ and $g(X)(X-r)$ (after expansion) are congruent modulo p. It is a stronger statement than $f(a) \equiv g(a)(a-r) \pmod{p}$ for all $a \in \mathbb{Z}$.

Theorem 8.6. *Suppose p is a prime and $f(X) \in \mathbb{Z}[X]$ has a leading coefficient not divisible by p. Then the equation $f(x) \equiv 0$ (mod p) has at most $\deg(f)$ solutions.*

Proof. We proceed by induction on $\deg(f)$. For the base case $\deg(f) = 1$, $f(X) = aX + b$, where $p \nmid a$. Then a is invertible modulo p, and we have

$$ax + b \equiv 0 \pmod{p} \iff x \equiv -ba^{-1} \pmod{p}$$

so the equation has a unique solution, and indeed $1 \leq \deg(f) = 1$.

Now assume the theorem holds for $\deg(f) \leq d$, for some $d \in \mathbb{N}$. Suppose $f(X) \in \mathbb{Z}[X]$ has degree $d+1$. If $f(X) \equiv 0 \pmod{p}$ has no solution, the statement is trivial. If $f(r) \equiv 0 \pmod{p}$, then by Proposition 8.4 we have $f(X) \equiv (X-r)g(X)$ for some $g(X) \in \mathbb{Z}[X]$ of degree d. Then

$$f(a) \equiv 0 \pmod{p} \iff (a-r)g(a) \equiv 0 \iff p \mid (a-r)g(a) \iff a - r \equiv 0 \text{ or } g(a) \equiv 0 \pmod{p}.$$

By the induction hypothesis $g(a) \equiv 0 \pmod{p}$ has at most $\deg(g) = d$ solutions. Therefore $f(a) \equiv 0$ has at most one more, i.e. $d + 1 = deg(f)$. This shows the theorem holds for $\deg(f) = d+1$. This proves the induction step, so the theorem holds for all values of $\deg(f)$. □

As a corollary, we obtain the following relationship between polynomials modulo p and the functions they define.

Corollary 8.7. *Suppose p is a prime, and $f(X),\ g(X) \in \mathbb{Z}[X]$ satisfy $\deg(f) = \deg(g) < p$. Then*

$$f(a) \equiv g(a) \ (\textit{mod}\ p) \textit{ for all } a \in \mathbb{Z} \iff f(X) \equiv g(X) \ (\textit{mod}\ p).$$

Proof. One direction is obvious: if $f(X) \equiv g(X) \pmod p$, then $f(a) \equiv g(a) \pmod p$ for all $a \in \mathbb{Z}$.

Assume $f(a) \equiv g(a) \pmod p$ for all $a \in \mathbb{Z}$. Let $h(X) = f(X) - g(X)$ and note $\deg(h) < p$. Let $\widetilde{h}(X)$ be obtained from $h(X)$ by removing the terms with coefficients that are $\equiv 0 \pmod p$. Then $\widetilde{h}(X) \equiv h(X) \pmod p$, so that $\widetilde{h}(a) \equiv f(a) - g(a) \equiv 0 \pmod p$. Then $\widetilde{h}(X)$ must be zero. Indeed otherwise $\widetilde{h}(X)$ has degree $< p$ and its leading coefficient is not divisible by p by construction, yet it has infinitely many roots, contradicting the theorem. Therefore since $f(X) - g(X) = h(X) \equiv \widetilde{h}(X) \pmod p$, we have $f(X) \equiv g(X) \pmod p$. □

As an application, we have the following.

Proposition 8.8. *For $k = 1, \cdots, p-1$, let*

$$S_k = \textit{ sum of products of distinct } k\textit{-subsets of } \{1, \cdots, p-1\}.$$

Then

(1) $S_k = \begin{cases} 0 \ (\textit{mod}\ p) & \textit{if } k \neq p-1 \\ -1 \ (\textit{mod}\ p) & \textit{if } k = p-1 \end{cases}$

(2) $S_{p-2} \equiv 0 \ (\textit{mod}\ p^2)$ *if* $p \geq 5$.

Proof. The polynomial $X^{p-1} - 1$ has $p-1$ roots $x = 1, \cdots, p-1$ by Fermat's little theorem. Therefore

$$X^{p-1} - 1 \equiv (X-1)(X-2)\cdots(X-p+1) \pmod p.$$

Expanding the product, we also have

$$(*) \qquad (X-1)(X-2)\cdots(X-p+1) = X^{p-1} - S_1X^{p-2} + S_2X^{p-3} - \cdots - S_{p-2}X + S_{p-1}.$$

Comparing the coefficients, (1) follows.

Now taking $X = p$, we have

$$(p-1)! = p^{p-1} - S_1p^{p-2} + S_2p^{p-3} + \cdots - S_{p-2}p + S_{p-1}.$$

Now $S_{p-1} = (p-1)!$ by definition of S_k, so

$$0 = p^{p-1} - S_1p^{p-2} + S_2p^{p-3} + \cdots + S_{p-3}p^2 - S_{p-2}p$$

hence

$$0 = p^{p-2} - S_1p^{p-3} + S_2p^{p-4} + \cdots + S_{p-3}p - S_{p-2}.$$

Then $S_{p-2} \equiv S_{p-3}p \pmod{p^2}$. Since $p \mid S_{p-3}$ by the first part, we obtain $p^2 \mid S_{p-2}$. Note we require $p \geq 5$ for S_{p-3} to exist. □

9. Multiplicative Order modulo m

Let $a, m \in \mathbb{N}$, $(a,m) = 1$. Recall that by Euler's theorem we have

$$a^{\varphi(m)} \equiv 1 \pmod m.$$

For instance, taking $m = 15$, $a = 2$, $a^{\varphi(15)} = 2^8 = 256 \equiv 1 \pmod{15}$. Now note that also

$$2^4 = 16 \equiv 1 \pmod{15}.$$

Then $\varphi(15) = 8$ is not the smallest exponent h for which $2^h \equiv 1 (\text{mod } 15)$.

Definition 9.1. Assume $(a, m) = 1$. The smallest positive integer h such that $a^h \equiv 1 \pmod m$ is called the **(multiplicative) order** of a modulo m, and denoted $\text{ord}_m(a)$.

For instance, $\text{ord}_{15}(2) = 4$. Note that $\text{ord}_{15}(2) \mid \varphi(15)$. In fact, we have:

Proposition 9.2. *Suppose $a, m \in \mathbb{N}$, $(a,m) = 1$, and $h = \mathrm{ord}_m(a)$. For $k \in \mathbb{Z}_{\geq 0}$, we have*

$$a^k \equiv 1 \ (mod\ m) \iff h \mid k.$$

Proof. It's clear that if $h \mid k$, then $a^k \equiv 1 \pmod n$. Conversely, suppose $a^k \equiv 1 \pmod n$. There exist unique $q, r \in \mathbb{Z}_{\geq 0}$, such that $k = qh + r$, $h \geq 0$, $h > r \geq 0$. Then

$$1 \equiv a^k = a^{qh+r} = (a^h)^q \cdot a^r \overset{m}{\equiv} 1^q \cdot a^r = a^r.$$

Now since $0 \leq r < h$, and h is the smallest positive integer such that $a^h \equiv 1 \pmod m$, we must have $r = 0$, i.e. $h \mid \varphi(m)$. □

Corollary 9.3. *For all $a, m \in \mathbb{N}$, $(a,m) = 1$, we have* $\mathrm{ord}_m(a) \mid \varphi(m)$.

Example: Suppose $(a, 15) = 1$. Since $\varphi(15) = 8$, $\mathrm{ord}_{15}(a)$ is one of $1, 2, 4$ or 8. Indeed, the invertible residues modulo 15 are

$$1,\ 2,\ 4,\ 7,\ 8,\ 11,\ 13,\ \text{and } 14,$$

having orders

$$1,\ 4,\ 2,\ 4,\ 4,\ 2,\ 4,\ \text{and } 2.$$

We can apply multiplicative orders to the problem of detecting primes.

Lemma 9.4. *Let p be prime, and $a \in \mathbb{N}$. If q is a prime divisor of $\frac{a^p-1}{a-1}$, then either $q = p$, or $q \equiv 1 (mod\ p)$.*

Proof. Suppose $a^p - 1 = k(a-1)$ and $q \mid k$. Then $a^p \equiv 1 \pmod q$. For $h = \mathrm{ord}_q(a)$, we have $h|p$, so either $h = 1$ or $h = p$.

If $h = p$, then $a^{q-1} \equiv 1 \pmod q$ implies $p \mid q - 1$, i.e. $q \equiv 1 \pmod p$.

If $h = 1$, then $a \equiv 1 \pmod q$, and

$$0 \equiv \frac{a^p - 1}{a - 1} = 1 + a + a^2 + \cdots a^{p-1} \equiv p \pmod q,$$

which implies $q \mid p$, hence $q = p$. □

Proposition 9.5. *Let p be prime. There are infinitely many primes congruent to 1 modulo p.*

Proof. We adapt Euclid's proof using the lemma. Suppose that $p_1, \cdots, p_k$ are distinct primes congruent to 1 modulo p. Let $a = p \cdot p_1 \cdots p_k$, and suppose q is a prime divisor of $\frac{a^p-1}{a-1}$. By the lemma, either $q = p$ or $q \equiv 1 (\mathrm{mod}\ p)$. By construction, $a^p \equiv 0 \not\equiv 1 \pmod p$, so $q \neq p$. On the other hand, $a^p \not\equiv 1 (\mathrm{mod}\ p_i)$, for $i = 1, \cdots, r$, so $q \notin \{p_1, \cdots, p_r\}$. Then q is a prime congruent to 1 modulo p which is not in the list $\{p_1, \cdots, p_r\}$. By induction (or contradiction), there must be infinitely many such primes. □

As before, this proof can be adapted to an algorithm that generates an infinite list of primes congruent to 1 modulo p.

Definition 9.6. A **Mersenne prime** is a prime of the form $2^p - 1$, for some prime p.

Note that if $2^n - 1$ is prime, n must be prime, because $2^a - 1$ divides $2^{ab} - 1$ in general. The following proposition restricts the possible prime divisors of $2^p - 1$. This can be used in searching for Mersenne primes, for instance.

Proposition 9.7. *Suppose q is a prime divisor of $2^p - 1$, and p is an odd prime. Then $q \equiv 1$ $(mod\ 2p)$.*

Proof. We previously showed that if $q \mid \frac{a^p-1}{a-1}$, then either $q = p$ or $q \equiv 1 \pmod p$. Take $a = 2$. If $p = q$, then $p \mid 2^p - 1$. But also $p \mid 2^{p-1} - 1$ by Fermat's Theorem, so $p \mid 2^p - 2^{p-1} = 2^{p-1}$. This can't happen because p is odd, therefore $q \equiv 1 (\mathrm{mod}\ p)$. Since q must be odd, we get $q \equiv 1 (\mathrm{mod}\ 2p)$. □

It is conjectured, but not known, that there are infinitely many Mersenne primes. A few dozen are known however, and these include some of the largest primes discovered.

Definition 9.8. The nth **Fermat number** is $F_n = 2^{2^n}+1$. A **Fermat prime** is a Fermat number that's prime.

The first five Fermat numbers $F_0, \cdots F_4$ are $2, 5, 17, 257, 65537$, which are all primes. In the 1600s, Fermat conjectured that *all* Fermat numbers are prime. However, as Euler discovered, F_5 is divisible by 641. In fact, as of 2014, $F_0, \cdots, F_4$ are the *only* known Fermat primes, and $F_5, \cdots, F_{32}$ have been verified as composite.

The following proposition restricts the possible prime divisors of a Fermat number $2^{2^n}+1$.

Proposition 9.9. *Suppose q is prime and $q \mid 2^{2^n}+1$. Then $q \equiv 1 \pmod{2^{n+1}}$.*

Proof. The assumption is $2^{2^n} \equiv -1 \pmod{q}$. Squaring both sides we get $2^{2^{n+1}} \equiv 1 \pmod{q}$. Then $\mathrm{ord}_q(2) \mid 2^{n+1}$, i.e. $\mathrm{ord}_q(2) = 2^k$ for some $k \le n+1$. If $k \le n$, then $2^k \mid 2^n$, which would imply $2^{2^n} \equiv 1 \pmod{q}$. This contradicts the assumption (since q is odd), therefore $k = n+1$ necessarily. Now $2^{q-1} \equiv 1 \pmod{q}$ by Fermat's theorem, so $\mathrm{ord}_q(2) = 2^{n+1} \mid q-1$, in other words $q \equiv 1 \pmod{2^{n+1}}$. □

The first five Fermat numbers $F_0, \cdots F_4$ are $2, 5, 17, 257, 65537$, which are all primes. In the 1600s, Fermat conjectured that *all* Fermat numbers are prime. However, as Euler discovered, 641 divides F_5. Fermat must have made a calculation mistake, since 641 is among the first few possibilities suggested by Proposition 9.9. As of 2014, $F_0, \cdots, F_4$ are the *only* known Fermat primes, and $F_5, \cdots, F_{32}$ have been verified as composite.

We now take a closer look at multiplicative orders modulo a number m. In general one has to do some computation to determine the order. On the other hand once $\mathrm{ord}_m(a)$ is known, it's easy to find $\mathrm{ord}_m(a^e)$ for $e \ge 1$.

Lemma 9.10. *Suppose $n = \mathrm{ord}_m(a)$ and $e \mid n$. Then $\mathrm{ord}_m(a^e) = \frac{n}{e}$.*

Proof. Let $b = a^e$ and $n = n'e$. We have

$$b^{n'} = (a^e)^{n'} = a^n \equiv 1 \pmod{m} \Longrightarrow \mathrm{ord}_m b \mid n'.$$

On the other hand

$$1 \equiv b^{\mathrm{ord}_m(b)} = a^{e\cdot \mathrm{ord}_m(b)} \pmod{m} \Longrightarrow n \mid e \cdot \mathrm{ord}_m(b) \Longrightarrow n' \mid \mathrm{ord}_m(b),$$

so $\mathrm{ord}_m(b) = n' = \frac{n}{e}$. □

Lemma 9.11. *Suppose $(a, m) = 1$, so that $\mathrm{ord}_m(a)$ exists. If $(e, n) = 1$, then $\mathrm{ord}_m(a^e) = \mathrm{ord}_m(a)$.*

Proof. Let $n = \mathrm{ord}_m(a)$ and $b = a^e$. Then

$$a^{e\cdot \mathrm{ord}_m(b)} = b^{\mathrm{ord}_m(b)} \equiv 1 \pmod{m} \Longrightarrow n \mid e \cdot \mathrm{ord}_m(b) \Longrightarrow n \mid \mathrm{ord}_m(b) \quad \text{(since } (e,n)=1\text{)}$$

At the same time

$$b^n = (a^n)^e \equiv 1 \pmod{m} \Longrightarrow \mathrm{ord}_m(b) \mid n,$$

therefore $\mathrm{ord}_m(b) = n$. □

Combining the two lemmas,

Proposition 9.12. *Assume $m > 1$, $(a, m) = 1$, and $n = \mathrm{ord}_m(a)$. For $e \ge 1$,*

$$\mathrm{ord}_m(a^e) = \frac{n}{(n, e)}$$

Proof. Write $e = ge'$, where $g = (n, e)$. Then $(e', n) = 1$, so $\mathrm{ord}_m(a^{e'}) = n$ by Lemma 9.11, and

$$\mathrm{ord}_m(a^e) = \mathrm{ord}_m((a^{e'})^g) = \frac{\mathrm{ord}_m(a^{e'})}{g} = \frac{n}{g}$$

by Lemma 9.10. □

Example: Let $m = 35$, and $a = 3$. Then $\varphi(m) = 35 = \varphi(5)\varphi(7) = 24$, so $\mathrm{ord}_{35}(3) \mid 24$. Then

$$\mathrm{ord}_{35}(3) \in \{1, 2, 4, 6, 8, 12, 24\}.$$

Now

$$3^2 \equiv 9,\ 3^3 \equiv 27,\ 3^4 \equiv 11 \pmod{35} \Longrightarrow 3^6 \equiv 99 \equiv 29 \bmod (m),$$

and

$$3^8 = 3^2 \cdot 3^6 \equiv 9 \cdot (-6) = -54 \equiv 16,$$

so $\mathrm{ord}_{35}(3) \neq 1, 2, 3, 4, 6, 8$. Since

$$3^{12} = 3^8 \cdot 3^4 \equiv 16 \cdot 11 = 176 \equiv 1 \pmod{35},$$

we have $\mathrm{ord}_{35}(3) = 12$. Then

$$\mathrm{ord}_{35}(9) = \mathrm{ord}_{35}(3^2) = \frac{12}{(12, 2)} = 6,$$

$$\mathrm{ord}_{35}(27) = \mathrm{ord}_{35}(3^3) = \frac{12}{(12, 3)} = 4,$$

$$\mathrm{ord}_{35}(11) = \mathrm{ord}_{35}(3^4) = \frac{12}{(12, 4)} = 3,$$

and since $3^5 \equiv 11 \cdot 3 = 33$,

$$\mathrm{ord}_{35}(33) = \frac{12}{(12, 5)} = 12.$$

Of the 24 numbers which are relatively prime to 35, 12 of them have the form 3^e.

10. Primitive Root Theorem

Let $m > 1$ and $(a, m) = 1$. Recall that by Euler's theorem $\mathrm{ord}_m(a) \mid \varphi(m)$. Usually there is no number a with $\mathrm{ord}_m(a) = \varphi(m)$. For instance if $m = 8$, then $\mathrm{ord}_m(a) = 2$, unless $a \equiv 1 \pmod 8$ (in which case $\mathrm{ord}_m(a) = 1$). On the other hand if $m = 5$ we have $\mathrm{ord}_m(3) = 4 = \varphi(5)$.

Definition 10.1. If $m > 1$, $(a, m) = 1$, and $\mathrm{ord}_m(a) = \varphi(m)$, then a is called a **primitive root modulo** m. In that case m is said to **admit (or have) a primitive root**.

We will prove the following theorem in several steps.

Theorem 10.2 (Primitive Root Theorem). *A modulus $m > 1$ admits a primitive root if and only if $m = 2, 4, p^\alpha$, or $2p^\alpha$, where p is an odd prime, $\alpha \geq 1$.*

First a simple observation regarding orders of products of elements.

Lemma 10.3. *Suppose $(a, m) = (b, m) = 1$, $\mathrm{ord}_m(a) = k$, $\mathrm{ord}_m(b) = h$. Then $\mathrm{ord}_m(ab) | \mathrm{lcm}(k, h)$. If $(k, h) = 1$, then $\mathrm{ord}_m(ab) = kh$.*

Proof. Let $l = lcm(k, h) = kh/g$. Writing $h = h'g$ and $k = k'g$, we have

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

which shows $\text{ord}_m(ab)|l$. Now suppose $(h, k) = 1$. Then $\text{lcm}(h, k) = hk$ so $\text{ord}_m(ab)|hk$. If $n = \text{ord}_m(ab)$, then $(ab)^n \equiv 1 \pmod m$, and

$$a^n \equiv b^{-n} \implies 1 \equiv a^{kn} \equiv b^{-kn} \implies b^{kn} \equiv 1 \implies h|kn \implies h|n.$$

Similarly $k|n$, and since $(h, k) = 1$, $hk|n$. This shows $hk = \text{ord}_m(ab)$.

□

Although in general $\text{ord}_m(ab)$ may be strictly smaller than $\text{ord}_m(a)\text{ord}_m(b)$, it so happens that there always exists *some* c with $\text{ord}_m(c) = \text{ord}_m(a)\text{ord}_m(b)$. To prove the primitive root theorem we will use a slightly more general version of this fact:

Lemma 10.4. *Suppose $a_1, \cdots, a_r$ are coprime to m, with $\text{ord}_m(a_i) = k_i$. If $k = \text{lcm}(k_1, \cdots, k_r)$, then there exists some a, $(a, m) = 1$, such that $\text{ord}_m(a) = k$.*

Proof. Suppose that $k = p_1^{\alpha_1} \cdots p_s^{\alpha_s}$ is the prime factorization of $k = \text{lcm}(k_i)$. Then for each j, $1 \le j \le s$, there exists some a_i such that $p_j^{\alpha_j} \| k_i$. In other words, for some i we have $k_i = p_j^{\alpha_j} k'_i$, where $p_j \nmid k'_i$. Now if we put $b_j = a_i^{k'_i}$, then $\text{ord}_m(b_j) = p_j^{\alpha_j}$ by Proposition 9.12.

This shows we can find $b_1, \cdots, b_s$ such that $\text{ord}_m(b_j) = p_j^{\alpha_j}$. Now $\text{lcm}_j\{\text{ord}_m(b_j)\}$ is the same as $k = \text{lcm}_i\{\text{ord}_m(a_i)\}$. So to find an element with order equal to k as in the statement, we can replace the numbers $\{a_i\}$ with $\{b_j\}$, and assume $k_i = \text{ord}_m(a_i)$ are all pairwise relatively prime.

We now proceed by induction on r. If $r = 1$ there is nothing to show. Suppose the proposition holds for $r < n$. Given $a_1, \cdots, a_{n+1}$, we assume their orders are pairwise relatively prime. By the induction hypothesis there exists some a' whose order modulo m is $k' = \text{lcm}(a_1, \cdots, a_n)$. But since $k_i = \text{ord}_m(a_i)$ are all pairwise relatively prime, $k' = k_1 \cdots k_n$. Then $\text{ord}_m(a') = k'$, $\text{ord}_m(a_{n+1}) = k_{n+1}$, and $(k_{n+1}, k') = 1$. It follows that $a = a'a_{n+1}$ has order $k'k_{n+1} = k$, as desired. By induction, the proposition holds for all $n \in \mathbb{N}$. □

Now we can prove the existence of primitive roots modulo a prime p, on the way towards the full statement of Theorem 10.2.

Proposition 10.5. *A prime modulus admits primitive roots.*

Proof. Let p be prime. Let $k_1, \cdots, k_{p-1}$ be the orders modulo p of all the non-zero residues $1, \cdots, p-1$. By Lemma 10.4, there exists some number r with $\text{ord}_p(a) = \text{lcm}(k_1, \cdots, k_{p-1})$. We claim that r is a primitive root modulo p.

Since $a^{p-1} \equiv 1 \pmod p$ for $a = 1, \cdots, p-1$, we have $k_i \mid p-1$ for each i, hence $k \mid p-1 \implies k \le p-1$. Then $f(X) = X^k - 1$ has for its roots modulo p all the $p-1$ different non-zero residues. But since $\deg(f) \le p-1$, f can have at most $\deg(f)$ roots, therefore $\deg(f) = p-1$. In particular $\text{ord}_p(r) = p-1$, and r is a primitive root modulo p. □

Example 10.6. Take $p = 17$, and let $S = \{1, 2, \cdots, 16\}$. The proposition says there is a number $r \in S$ such that

$$S = \{1, r, r^2, \cdots, r^{16}\}.$$

We have $2^4 \equiv -1 (\text{mod } 17) \implies 2^8 \equiv 1$, so $\text{ord}_{17}(2) \le 8 \ne 16$, hence 2 is not a primitive root.

On the other hand, the numbers

$$3, 3^2, 3^3, \cdots, 3^{16}$$

are in order congruent to

$$3, 9, 10, 13, 5, 15, 11, 16, 14, 8, 7, 4, 12, 2, 6, 1.$$

In other words 3 is a primitive root modulo 17.

Theorem 10.2, which we are in the process of proving, says the moduli m that admit primitive roots are exactly $2, 4, p^\alpha$, and $2p^\alpha$ for odd primes p, $\alpha \geq 1$. The cases of 2 and 4 are trivial, and we showed primes p admit primitive roots.

Proposition 10.7. *If r is a primitive root for an odd prime p, then either r or $r + p$ is a primitive root for p^2.*

Proof. Let $m = \operatorname{ord}_{p^2}(r)$. From $r^{\varphi(p^2)} \equiv 1 (\text{mod } p^2)$, we know $m \mid \varphi(p^2) = p(p-1)$. On the other hand, from $r^m \equiv 1 \pmod{p^2} \Longrightarrow r^m \equiv 1 \pmod p$, we have $\operatorname{ord}_p(r) = p - 1 \mid m$. So:

$$m \mid p(p-1) \text{ and } p-1 \mid m,$$

which since p is prime implies

$$m = p-1 \text{ or } m = p(p-1).$$

For $s = r + p$, and $n = \operatorname{ord}_{p^2}(s)$, since $s \equiv r \pmod p$, for the same reason we havea

$$n = p-1 \text{ or } n = p(p-1).$$

Now

$$s^p = (r+p)^p = r^p + \binom{p}{1} r^{p-1} p + \cdots + \binom{p}{p} p^p \equiv r^p \pmod{p^2}.$$

Then if $m = p - 1$,

$$r^p \equiv r \pmod{p^2} \Longrightarrow s^p \equiv r^p \equiv r \not\equiv s \pmod{p^2} \Longrightarrow n \neq p-1 \Longrightarrow n = p(p-1).$$

In other words, if not r, then s is a primitive root for p^2. □

Example 10.8. We know 3 is a primitive root for $p = 17$. By the proposition, either 3 or $3+17 = 20$ must be a primitive root for $p^2 = 289$. In fact, both 3 and 20 are primitive roots for 289, and there are many others, as we will see.

Proposition 10.9. *If r is a primitive root modulo p^2, for p an odd prime, it is also a primitive root modulo p^α, for all $\alpha \geq 2$.*

Proof. We proceed by induction. The base case is given, so assume r is a primitive root for p^k, $2 \leq k \leq \alpha$. We show it is also a primitive root for $p^{\alpha+1}$.

Let $n = \operatorname{ord}_{p^{\alpha+1}}(r)$, so that $n \mid \varphi(p^{\alpha+1})$. Also

$$r^n \equiv 1 \pmod{p^{\alpha+1}} \Longrightarrow r^n \equiv 1 \pmod{p^\alpha} \Longrightarrow \varphi(p^\alpha) \mid n.$$

Since $\varphi(p^{\alpha+1}) = p^\alpha(p-1) = p\varphi(p^\alpha)$, we have either

$$\operatorname{ord}_{p^{\alpha+1}}(r) = \varphi(p^{\alpha+1}) \text{ or } \operatorname{ord}_{p^{\alpha+1}}(r) = \varphi(p^\alpha).$$

We claim $\operatorname{ord}_{p^{\alpha+1}}(r) \neq \varphi(p^\alpha)$, which is to say $r^{\varphi(p^\alpha)} \not\equiv 1 \pmod{p^{\alpha+1}}$.

From $r^{\varphi(p^{\alpha-1})} \equiv 1 \pmod{p^{\alpha-1}}$ we have

$$r^{\varphi(p^{\alpha-1})} = 1 + ap^{\alpha-1}$$

for some $a \in \mathbb{Z}$. Then

$$r^{\varphi(p^\alpha)} = (r^{\varphi(p^{\alpha-1})})^p = (1 + ap^{\alpha-1})^p = 1 + \binom{p}{1} ap^{\alpha-1} + \binom{p}{2} a^2 p^{2\alpha-2} + \cdots + \binom{p}{p} a^p p^{\alpha p - p}$$
$$\equiv 1 + ap^\alpha \pmod{p^{\alpha+1}}.$$

Then $r^{\varphi(p^\alpha)} \equiv 1 \pmod{p^{\alpha+1}}$ if and only if $p \mid a$. On the other hand since r is a primitive root for p^α, and $\varphi(p^{\alpha-1}) < \varphi(p^\alpha)$,

$$r^{\varphi(p^{\alpha-1})} = 1 + ap^{\alpha-1} \not\equiv 1 \pmod{p^\alpha} \Longrightarrow p \nmid a.$$

Therefore $r^{\varphi(p^\alpha)} \not\equiv 1 \pmod{p^{\alpha+1}}$, so $\operatorname{ord}_{p^{\alpha+1}}(r) \neq \varphi(p^\alpha)$, hence $\operatorname{ord}_{p^{\alpha+1}}(r) = \varphi(p^{\alpha+1})$. □

If r is a primitive root modulo m, then $\{r, r^2, \cdots, r^{\varphi(m)}\}$ are exactly the distinct non-zero residue classes modulo m. In other words, any integer a, with $1 \leq a < m$, $(a, m) = 1$ can be written as r^e for a unique $e \leq \varphi(m)$.

Example 10.10. 3 is a primitive root for 17 as well as for 17^2, hence also for all 17^α by the proposition. Let us take for instance $\alpha = 6$. We have $(5, 17^6) = 1$, so there is a unique positive integer $e \leq \varphi(17^6) = 17^5 \cdot 16$ such that $3^e \equiv 5 \pmod{17^6}$. A brief computer search reveals that $e = 17269781$. The number $3^{17269781}$ has well over eight million digits, and is the smallest power of 3 congruent to 5 modulo 17^6.

Primitive roots, when they do exist, are never unique.

Proposition 10.11. *If m admits a primitive root r, it admits exactly $\varphi(\varphi(m))$ primitive roots in total, corresponding to r^e, for $1 \leq e \leq \varphi(m)$, $(e, \varphi(m)) = 1$.*

Proof. This follows immediately from Proposition 9.12. Indeed

$$\mathrm{ord}_m(r^e) = \frac{\mathrm{ord}_m(r)}{(e, \mathrm{ord}_m(r))} = \frac{\varphi(m)}{(e, \varphi(m))},$$

so r^e is a primitive root if and only if $(e, \varphi(m)) = 1$. □

Example 10.12. From the previous example, $5 \equiv 3^e \pmod{17^6}$, for $e = 17269781$. Then

$$\mathrm{ord}_{17^6}(5) = \frac{\varphi(17^6)}{(e, \varphi(17^6))}.$$

Now $\varphi(17^6) = 17^5 \cdot 16$. Since e is odd, $(e, \varphi(17^6)) = (e, 17^5)$. It's easy to check that $17 \nmid e$, therefore $\mathrm{ord}_{17^6}(5) = \varphi(17^6)$. In particular, 5 is itself a primitive root modulo 17^6, hence also a primitive root modulo all other powers of 17.

Primitive roots modulo $2p^\alpha$ can be easily constructed from those modulo p^α.

Proposition 10.13. *Suppose r is a primitive root modulo p^α, for $\alpha \geq 1$, p an odd prime. Then either r or $r + p^\alpha$ is a primitive root for $2p^\alpha$.*

Proof. Let $s = r$ if r is odd, and $r + p^\alpha$ if r is even. Then in either case s is an odd primitive root modulo p^α. Let $n = \mathrm{ord}_{2p^\alpha}(s)$, which exists since $(s, 2p^\alpha) = 1$. Then arguing as before,

$$s^n \equiv 1 \pmod{2p^\alpha} \Longrightarrow s^n \equiv 1 (\mathrm{mod}\ p^\alpha) \Longrightarrow \varphi(p^\alpha) \mid n.$$

On the other hand

$$s^{\varphi(p^\alpha)} = s^{\varphi(2p^\alpha)} \equiv 1 \bmod (2p^\alpha) \Longrightarrow n \mid \varphi(2p^\alpha) = \varphi(p^\alpha),$$

hence $n = \varphi(p^\alpha) = \varphi(2p^\alpha)$. This shows s is a primitive root modulo $2p^\alpha$. □

We have shown that for odd primes p, and $\alpha \geq 1$, p^α and $2p^\alpha$ admit primitive roots. The following proposition completes the proof of Theorem 10.2.

Proposition 10.14. *Suppose $m > 1$ is a modulus admitting a primitive root. Then $m = 2$, 4, p^α, or $2p^\alpha$ for an odd prime p, $\alpha \geq 1$.*

Proof. Suppose that $m = ab$ with $(a, b) = 1$. If r is a primitive root modulo m, we have

$$r^{\varphi(m)} \equiv 1 \pmod{m} \Longrightarrow r^{\varphi(m)} \equiv 1 \pmod{a} \Longrightarrow \varphi(m) \mid \mathrm{ord}_a(r).$$

But since r is a primitive root modulo m and $a \mid m$, it must also be a primitive root modulo a, so $\mathrm{ord}_a(r) = \varphi(a)$. The same argument shows $\varphi(m) \mid \mathrm{ord}_b(r)$. It follows that $\varphi(m)$ divides $l = \mathrm{lcm}(\mathrm{ord}_a(r), \mathrm{ord}_b(r)) = \mathrm{lcm}(\varphi(a), \varphi(b))$. Since $\varphi(m) = \varphi(a)\varphi(b)$ and $\mathrm{lcm}(\varphi(a), \varphi(b)) =$

$\varphi(a)\varphi(b)/(\varphi(a),\varphi(b))$, that can only happen if $(\varphi(a),\varphi(b)) = 1$. Then necessarily either $a = 2$ or $b = 2$, since $\varphi(n)$ is even for $n > 2$.

We have shown that if m can be written as $m = ab$ with $(a, b) = 1$ and $a, b > 1$, then $a = 2$ and $b = p^\alpha$ for some odd prime p, or the other way around. Hence $m = 2p^\alpha$.

If $m = ab$ with $(a, b) = 1$ is not possible, then $m = p^\alpha$ for some prime p, $\alpha \geq 1$. If $p = 2$, then $\alpha = 1$ or 2, since $m = 2^3$ does not admit a primitive root. Then in this case $m = 2$, 4, or p^α for some odd prime p. □

11. Monday, May 1

With knowledge of primitive roots, we can now revisit an earlier problem. Recall that for $m > 1$, we have $(m-1)! \equiv -1 \pmod{m}$ if and only if m is prime, by Wilson's Theorem (Theorem 5.12). The following generalization is due to Gauss.

Theorem 11.1 (Gauss-Wilson)**.** *For $m > 1$,*

$$\prod_{\substack{1\leq a\leq m\\(a,m)=1}} a = \begin{cases} -1 \ (\mathit{mod}\ m) & \mathit{if}\ m = 2,\ 4,\ p^\alpha,\ \mathit{or}\ 2p^\alpha\ \mathit{for\ an\ odd\ prime}\ p \\ \ \ 1 \ (\mathit{mod}\ m) & \mathit{otherwise} \end{cases}$$

Proof. Let $S_m = \{a : 1 \leq a \leq m, (a, m) = 1\}$ be the set over which the product is taken. We can partition S_m into two sets

$$T_m = \{a \in S_m : a^2 \equiv 1 \pmod{m}\}, \quad T'_m = \{a \in S_m : a^2 \not\equiv 1 \pmod{m}\}.$$

If $a \in T'_m$, then $a^{-1} \not\equiv a \pmod{m}$, so that

$$\prod_{a\in T'_m} a \equiv 1 \pmod{m},$$

since each $a \in T'_m$ will cancel out with its inverse. Therefore

$$\prod_{a\in S_m} a \equiv \prod_{a\in T_m} a \pmod{m}.$$

Now consider the equation $X^2 \equiv 1 \pmod{m}$. If $m = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$, each $a \in T_m$ corresponds by the Chinese Remainder Theorem to a set of residues $a_1, \cdots, a_r$, modulo $p_i^{\alpha_i}$, $i = 1, \cdots r$, where $a_k^2 \equiv 1 \pmod{p_k^{\alpha_k}}$.

Modulo an odd prime p, the polynomial $f(X) = X^2 - 1$ has 2 distinct roots, and $f'(X) = 2X \not\equiv 0$. Therefore by Hensel's Lemma, there are exactly two solutions $\pm a$, to each $X^2 \equiv 1 \pmod{p^\alpha}$.

Now if there are two distinct odd primes p and q in the prime factorization of m with exponents α and β, for every $a \pmod{m}$ such that $a^2 \equiv 1 \pmod{m}$, there are three other distinct residues b, c, d such that

$$\begin{aligned} &b \equiv -a \pmod{p^\alpha}, && b \equiv -a (\mathrm{mod}\ q^\beta), && b \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq p \text{ or } q \\ &c \equiv +a \pmod{p^\alpha}, && c \equiv -a (\mathrm{mod}\ q^\beta), && c \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq p \text{ or } q \\ &d \equiv -a \pmod{p^\alpha}, && d \equiv +a (\mathrm{mod}\ q^\beta), && d \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq p \text{ or } q. \end{aligned}$$

Then $abcd \equiv 1$ $(p_i^{\alpha_i})$ for $i = 1, \cdots r$, hence $abcd \equiv 1 \pmod{m}$. It follows that $\prod_{a\in S_m} a \equiv \prod_{a\in T_m} a \equiv 1 \pmod{m}$ if m has two distinct odd primes.

Now suppose 2^α exactly divides m. If $\alpha > 2$ and c is an odd number that's a root of $X^2 - 1 \equiv 0 \pmod{2^\alpha}$, then $c^{\alpha-1} - c$, $2^{\alpha-1} + c$ and $-c$ are all distinct residues modulo 2^α. It follows that for every residue class a such that $a^2 \equiv 1 \pmod{m}$, there are distinct residue classes b, c, d such that

$$\begin{aligned} &b \equiv -a \pmod{2^\alpha}, && b \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq 2 \\ &c \equiv 2^{\alpha-1} - a \pmod{p^\alpha}, && c \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq 2 \\ &d \equiv 2^{\alpha-1} + a \pmod{p^\alpha}, && d \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq 2. \end{aligned}$$

Then we have $abcd \equiv (-a^2) \cdot (2^\alpha - a^2) \equiv a^4 \equiv 1 \pmod{2^\alpha}$, and also $abcd \equiv a^4 \equiv 1 \pmod{p_i^{\alpha_i}}$.

If 2^2 occurs in the prime factorization of m, along with p^α for some odd prime p, then again for each residue a mod m satisfying $a^2 \equiv 1 \pmod{m}$ we have distinct residues b, c, d such

$$\begin{aligned} &b \equiv a \pmod{4}, && b \equiv -a (\text{mod } p^\alpha), && b \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq 2 \text{ or } q \\ &c \equiv a + 2 \pmod{4}, && c \equiv +a (\text{mod } p^\alpha), && c \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq 2 \text{ or } q \\ &d \equiv a + 2 \pmod{4}, && d \equiv -a (\text{mod } p^\alpha), && d \equiv a \pmod{p_i^{\alpha_i}} \text{ for } p_i \neq 2 \text{ or } q. \end{aligned}$$

Then $abcd \equiv a^2(a+2)^2 \equiv 1 \pmod{4}$, and $abcd \equiv a^4 \equiv 1 (\text{mod } p_i)$ for all odd primes $p_i \mid m$. Once again it follows that $\prod_{a \in S_m} a = \prod_{a \in T_m} a = 1$. We have therefore proved that if $\prod_{a \in S_m} a \not\equiv 1 \pmod{m}$, then necessarily $m = 2$, 4, p^α, or $2p^\alpha$, where p is an odd prime. Then by Theorem 10.2, m admits a primitive root.

Suppose there exists r such that $S_m = \{r, r^2, \cdots, r^{\varphi(m)}\}$. Then

$$\prod_{a \in S_m} a \equiv r^{1+2+\cdots+\varphi(m)} \equiv r^{\frac{\varphi(m)(\varphi(m)+1)}{2}} \pmod{m}.$$

If $\varphi(m) = 1$, i.e. $m = 2$, the above product is simply $r^1 \equiv -1 (\text{mod } 2)$ so the theorem holds. Otherwise $\varphi(m) = 2k$, so that

$$\prod_{a \in S_m} a \equiv r^{k(\varphi(m)+1)} \equiv (r^{\varphi(m)})^k \cdot r^k \equiv r^k \pmod{m}$$

by Euler's Theorem. We claim $r^k \equiv -1 \pmod{m}$. Since r is a primitive root and the class -1 belongs to S_m, we must have $r^l \equiv -1 \pmod{m}$ for some l. Then

$$r^{2l} \equiv 1 \Longrightarrow \varphi(m) \mid 2l \Longrightarrow k \mid l \Longrightarrow l = k \text{ or } 2k,$$

since $1 \leq l \leq \varphi(m)$. But $r^{2k} \equiv 1 \not\equiv -1$, so $l = k$ and $r^k \equiv -1$. This shows that if $m = 2$, 4, p^α or $2p^\alpha$, then $\prod_{a \in S_m} a \equiv -1 \pmod{m}$, finishing the proof of the theorem. □

12. Cryptography: Introduction

Cryptography involves transforming a message, the *plaintext*, into a disguised message, the *ciphertext*, and transmitting it in such a way that (ideally) only the intended recipient of the message can recover the plaintext. The plaintext does not have to be actual text, it can be any information we wish to conceal. *Encryption* is the transformation of plaintext into ciphertext, and *decryption* is the reverse process. A cryptosystem is a set of procedures for carrying this out.

Number theoretic cryptosystems often consider the plaintext as a number modulo m for some (large) m, and apply some invertible procedure using modular arithmetic.

Affine Cryptosystems

An affine cryptosystem is a simple number theoretic cryptosystem involving *alphabet substitution.* Suppose a message x is written using an alphabet with N letters. For instance ordinary text using only capital letters $A, \cdots, Z$, plus blank space, can be written with an alphabet of $N = 27$ letters.

To apply an affine cipher,

(1) Label the alphabet with $0, \cdots N-1$, e.g.

$$A = 0,\ B = 1,\ C = 2\ , \cdots,\ Z = 25,\ \text{“ ”} = 26.$$

(2) Pick k: size of a message block, and split the message into chunks of k letters, for instance if $k = 4$, we split up

HELLO THIS IS A CODED MESSAGE

into

"HELL", "O TH", "IS I", "S A ", "CODE", "ED M", "ESSA", "GE "

where we've padded the last block with two blanks so it contains $k = 4$ letters.

(3) To each block, assign a numerical equivalent mod N^k. For instance if a k-block has letters corresonding to numbers $a_0, \cdots, a_{k-1}$ with $0 \leq a_i \leq N-1$, assign to it the number

$$t = a_{k-1}N^{k-1} + a_{k-2}N^{k-2} + \cdots + a_1N + a_0 \pmod{N^k}.$$

Note that x_i can be recovered from t, since

$$a_0 \equiv t \pmod{N},$$

$$a_1 \equiv \frac{1}{N}(t - a_0) \pmod{N},$$

$$a_2 \equiv \frac{1}{N^2}(t - a_0 - a_1N) \pmod{N},$$

etc.

(4) Choose $a, b \in \{1, \cdots, N^k\}$, such that $(a, N) = 1$, and apply

$$\psi(x) = ax + b \pmod{N^k}$$

to t.

(5) Turn each $u = \psi(t) \pmod{N^k}$ into a message block, using the alphabet.

Affine cryptosystems, similar to alphabet substitution, are susceptible to a letter frequency attack. Note that since $t \equiv a_0 \pmod{N}$ above, we have $\psi(t) \equiv aa_0 + b \pmod{N}$. Then a_0 from each letter block is in fact encrypted with a simple substitution scheme. If the message is long enough, a_0 may be guessed based on frequency of letters in the language of the plaintext (if known). Once the a_0 from each block is found, then a_1 may be guessed based on frequency, etc.

The information required to encrypt or decrypt a message is called a *key*. In an *asymmetric* cryptosystem different keys are used for encrypting and decrypting. These are typically called the *public* and *private* keys, respectively.

RSA (named after its pioneers Rivest, Shamir, and Adleman), is one of the mostly commonly used asymmetric cryptosystems today. It's described as follows.

Suppose that Bob wishes to send Alice a secret message using RSA.

Step 0: Alice generates a public/private key pair

(1) Alice picks two large primes p and q at random. Typically each a few hundred digits in length.
(2) She computes $n = pq$.
(3) She computes $\ell = \mathrm{lcm}(p-1, q-1)$.
(4) She chooses an integer $e \in \{1, \cdots, \ell\}$ such that $(e, \ell) = 1$, at random. The pair (e, n) is Alice's *public key*. This she announces.
(5) She computes $d \in \{1, \cdots, \ell\}$ such that $d \cdot e \equiv 1 \pmod{\ell}$. The pair (n, d) is Alice's *private key*, which she keeps a secret.

Alice may use the same public/private key pair for communicating with several correspondents. Thus Step 0 may already have been carried out before Bob has the intention to communicate.

Remark 12.1. Once n,e and d are computed, the algorithm does not require p, q, or ℓ. But Alice must clearly keep them secret, since the private key (n, d) can be computed from them.

Step 1: Bob encrypts and transmits the message using Alice's public key

(1) Bob reads Alice's public key pair (n, e).
(2) By a method agreed beforehand, Bob splits the message M into smaller chunks $M_1, \cdots, M_k$ of equal size.
(3) He converts each M_i to a number $m_i \in \{0, \cdots n-1\}$.
(4) For each $m_i = m_{\text{plain}}$, Bob computes $c_i = m_{\text{cipher}} \in \{0, \cdots, n-1\}$, via
$$m_{\text{cipher}} \equiv m_{\text{plain}}^e \pmod n.$$
(5) Bob transmits c_i to Alice, for each $i = 1, \cdots, k$.

Step 2: Alice decrypts the cipher from Bob using her private key

To recover the plaintext from $c_1, \cdots, c_k$:

(1) For each $m_{\text{cipher}} = c_i$, Alice computes $m \in \{0, \cdots, n-1\}$ where
$$m \equiv m_{\text{cipher}}^d (\text{mod } n).$$
Then in fact $m = m_{plain} = m_i$. (See below for explanation)
(2) Alice recovers the plaintext chunks M_i from the numbers m_i (by the method agreed beforehand).
(3) She recovers the message M from the chunks M_i.

Why it works

RSA revolves around the fact that with n, d, e as above, for $x \in \{0, \ldots, n-1\}$, we have
$$x^{de} \equiv x \pmod n.$$
Indeed, suppose $\ell = (p-1)k = (q-1)m$. As $de \equiv 1 \pmod \ell$, we can write $de = 1 + r\ell$ for some r. Then by Fermat's little theorem,
$$x^{de} = x^{r\ell+1} = (x^{p-1})^{rk} \cdot x \equiv x \pmod p.$$
Similarly
$$x^{de} = x^{r\ell+1} = (x^{q-1})^{rm} \cdot x \equiv x \pmod q,$$
therefore
$$x^{de} \equiv x \pmod n$$
by the uniqueness part of the Chinese Remainder Theorem.

Remark 12.2. Note that $\varphi(n) = (p-1)(q-1)$ may be used instead of ℓ. Since $\ell \mid \varphi(n)$, $de \equiv 1 \pmod{\varphi(n)}$ implies $de \equiv 1 \pmod \ell$, this changes very little. This is in fact the original RSA. However, it is a bit more efficient to work with $\ell = \text{lcm}(p-1, q-1)$ instead, since it will be smaller than $\varphi(n)$.

Why RSA is (relatively) secure

The idea is that it is easy to generate large prime numbers p, q, and to compute $\ell = \text{lcm}(p-1, q-1)$, as well as the inverse of a number e modulo ℓ. But it is difficult to factor a large integer n, even knowing that it is of the form pq for two primes of roughly equal size.

No *polynomial-time* algorithm is known for factoring large numbers. That is to say, if $d = d(n)$ is the number of digits of n (in some fixed base), then no algorithm is known that can factor n in time $t = t(n)$ that is proportional to a polynomial function of d. It is expected that no such algorithm exists. However, it is also expected that the problem of factoring large numbers is easier than the most difficult non-polynomial time problems.

Much of the encrypted transmission over the internet relies on RSA for security, at least as a first step. If a highly efficient algorithm is ever found that can quickly factor large numbers into primes, we will have need of a replacement. Though it is unlikely such an algorithm exists, "quantum computers" can indeed factor large numbers in polynomial time.

Diffie-Hellman Key Exchange

With *symmetric* encryption schemes, the same key used to encrypt plaintext is also used for decrypting ciphertext. Alice and Bob must then agree on a shared secret key *before* they can begin communicating securely. Rather than for one party to transmit the shared secret in an insecure manner, it is possible for Alice and Bob to construct the same shared secret simultaneously.

The Diffie-Hellman Key Exchange is one such method. To begin with, first Alice (say) decides which parameters to choose.

Step 0: Alice and Bob agree on parameters

(1) Alice picks a large prime p at random.
(2) She computes a primitive root g modulo p.
(3) She transmits the parameters (p, g) to Bob (in public).

Then the shared secret is constructed as follows.

Step 1: Alice constructs a private/public key pair

(1) Alice chooses a random number $k_A \in \{1, \cdots, p-1\}$. This is her private key.
(2) She computes $K_A \in \{1, \cdots, p-1\}$ where
$$K_A \equiv g^{k_A} \pmod{p}.$$
This is her public key.
(3) She transmits K_A to Bob (in public).

Step 2: Bob constructs a key pair, and computes the shared secret

(1) Bob receives the parameters p, g and Alice's public key K_A.
(2) He chooses a private key: a random number $k_B \in \{1, \cdots, p-1\}$.
(3) He computes the public key $K_B \in \{1, \cdots, p-1\}$ where
$$K_B \equiv g^{k_B} \pmod{p}.$$
(4) The secret $s_B \in \{1, \cdots, p-1\}$ is computed by
$$s_B \equiv K_A^{k_B} \pmod{p}.$$
(5) Bob transmits K_B to Alice (in public).

Step 3: Alice computes the shared secret

(1) Alice receives Bob's public key K_B.
(2) She computes $s_A \in \{1, \cdots, n-1\}$ satisfying
$$s_A \equiv K_B^{k_A} \pmod{p}.$$

Now
$$s_A \equiv k_B^{k_A} \equiv (g^{k_B})^{k_A} = g^{k_B \cdot k_A} = (g^{k_A})^{k_B} \equiv k_A^{k_B} \equiv s_B \pmod{p}.$$

Thus Alice and Bob possess the same shared secret $s = s_A = s_B$. They may now pick a suitable symmetric encryption scheme and communicate back and forth, using the shared secret s to both encrypt and decrypt messages.

The security of the Diffie-Hellman key exchange is based on the difficulty of solving the equation

$$K_A \equiv g^{k_A} \pmod{p},$$

for k_A, while knowing p, g, and K_A (or the analogous one for k_B). This is called a *discrete logarithm* computation. Problems of this type are the basis for numerous encryption schemes.

The Diffie-Hellman Key Exchange relied on the following computation

$$K_B^{k_A} \equiv (g^{k_B})^{k_A} = g^{k_B \cdot k_A} = g^{k_A \cdot k_B} = (g^{k_A})^{k_B} \equiv K_A^{k_B} \pmod{p},$$

where (k_A, K_A), (k_B, K_B) are Alice and Bob's private/public key pairs, and g is an agreed-upon primitive root modulo a large prime p. A fundamental fact that makes this exchange possible is the commutativity of multiplication:

$$k_A k_B = k_B k_A.$$

The same principle underlies another scheme that allows Alice and Bob to exchange messages without sharing keys.

The three-pass protocol

Suppose (s, t) is a pair of keys for an asymmetric encryption scheme using two functions E and D to encrypt and decrypt. In other words

$$m_{\text{cipher}} = E_s(m_{\text{plain}}), \quad m_{\text{plain}} = D_t(m_{\text{cipher}}).$$

Suppose furthermore that for any other key pair (u, v), we have the commutativity relation

$$D_t(E_u((m)) = E_u(D_t(m)).$$

Then it is possible for Alice to send an encrypted message to Bob without the two of them ever sharing keys. It would go as follows:

(1) Alice generates her own key pair (s, t).
(2) She encrypts the message $m = m_{\text{plain}}$ using the key s:
$$m_1 = E_s(m).$$
(3) Alice transmits m_1 to Bob. (first pass)
(4) Bob generates his own key pair (u, v).
(5) Bob encrypts Alice's encrypted message with his key u:
$$m_2 = E_u(m_1).$$
(6) Bob transmits m_2 back to Alice. (second pass)
(7) Alice applies her decryption key to Bob's message:
$$m_3 = D_t(m_2).$$
(8) Alice transmits m_3 back to Bob. (third pass)
(9) Bob now decrypts with his own key:
$$m' = D_v(m_3).$$

Indeed

$$\begin{aligned} m' &= D_v(D_t(E_u(E_s(m)))) \\ &= D_v(E_u(D_t(E_s(m)))) \quad \text{(since by assumption } D_t E_u = E_u D_t) \\ &= D_t(E_s(m)) = m. \end{aligned}$$

A particular implementation using modular arithmetic to define E and D is as follows.

Massey-Omura cryptosystem

First Alice and Bob agree on a large prime p and a primitive root g modulo p.

A key pair (s,t) consists of integers in $\{1 \le a \le p-1 : (a, p-1) = 1\}$ such that $st \equiv 1 \pmod{p-1}$. For numbers $m \in \{1, \cdots, p-1\}$ representing message blocks, set $E_s(m)$ and $D_t(m)$ to be numbers in $\{1, \cdots, p-1\}$ such that

$$E_s(m) \equiv m^s \pmod{p}, \quad D_t(m) \equiv m^t \pmod{p}.$$

If $de = k(p-1)+1$, we have

$$D_t(E_s(m)) = m^{st} = m^{k(p-1)+1} \equiv m \pmod{p}.$$

Now note that modulo p,

$$D_t(E_u(m)) \equiv (m^u)^t \equiv (m^t)^u \equiv E_u(D_t(m)).$$

Therefore the three-pass protocol can be implemented using this encryption system. Making the 9 mentioned steps concrete in this case is left as an exercise.

Another cryptosystem based on discrete logs requires the sender of the message to produce temporary keys, also called *ephemeral* keys, for every transmission. An example is as follows.

ElGamal

Suppose that Alice wants to receive message from Bob. She first generates a public/private key as follows.

Step 0: Alice generates and announces parameters, public key

(1) Alice chooses a large prime p, and computes a primitive root g.
(2) She then chooses a random number $a \in \{1, \cdots, p-1\}$ as a private key.
(3) Alice computes the public key $A = g^a \pmod{p}$.
(4) She announces the parameters (p, g), as well as her public key A.

Now suppose Bob wishes to send a message M to Alice.

Step 1: Bob prepares to transmit to Alice

(1) Bob receives the parameters (p, g) as well as Alice's public key A.
(2) He splits M up into chunks $M_1, \cdots M_k$, and converts each chunk M_i into a number $m_k \in \{1, \cdots, p-1\}$.

Now for *every* $i = 1, \cdots, k$, Bob peforms the following step.

Step 2: Bob transmits a cipher block

(1) To transmit a message m, first Bob chooses a random number $b \in \{1, \cdots, p-1\}$. This is the ephemeral key, to be thrown away after this transmission.
(2) Bob uses the ephemeral key b to compute b', $c' \in \{1, \cdots, p-1\}$ such that

$$b' \equiv g^b \pmod{p}, \quad c' \equiv A^b \equiv g^{ab} \pmod{p}.$$

(3) He encrypts the message block m by computing $c \in \{1, \cdots, p-1\}$ such that

$$c \equiv c' \cdot m \pmod{p}.$$

(4) He transmits (b', c) to Alice.

For the next message block Bob will begin again at Step 2, computing a new ephemeral key b, then B, C, etc. Meanwhile, Alice decrypts each transmission.

Step 3: Alice receives and decrypts a cipher block

(1) Alice receives the disguised key b' and the cipher c from Bob, where

$$b' \equiv g^b \pmod{p}, \quad c \equiv g^{ab} m \pmod{p}.$$

(2) Using her private key a, and B, she computes

$$B^a \equiv g^{ab} \pmod{p}.$$

(3) She computes the inverse h of g^{ab} modulo p, and decrypts m via

$$m \equiv hg^{ab}m \equiv hc \pmod{p}.$$

13. Monday, May 8

It's helpful at this point to introduce some basic concepts from group theory. These will be the most elementary aspects, and not mentioned again beyond this lecture.

Definition 13.1. A *group* is a set G, with a binary operation $\cdot : G \times G \to G$, that satisfies the following properties.

1. Associativity: $(g \cdot h) \cdot k = g \cdot (h \cdot k)$, for all $g, h, k \in G$.
2. Existence of Identity: there exists some $e \in G$, such that $e \cdot g = g \cdot e = g$ for all $g \in G$.
3. Existence of Inverses: for all $g \in G$, there exists some $g' \in G$, such that $g \cdot g' = g' \cdot g = e$.

The operation $\cdot$ is called *group multiplication.* In this generality its relation to multiplication of numbers is just metaphorical.

Example 13.2. Some examples:

1. The non-zero real numbers $\mathbb{R}^\times$, with ordinary multiplication for $\cdot$, define a group. The identity is 1, and the inverse of $r \in \mathbb{R}^\times$, is $\frac{1}{r}$.
2. The real numbers $\mathbb{R}$, together with addition, form a group. The identity is 0, and the inverse of $r \in \mathbb{R}$ is $-r$.
3. The integers $\mathbb{Z}$, together with addition, form a *subgroup* of $(\mathbb{R}, +)$.
4. For each $m > 1$, the classes of residues modulo m form a group under addition. The identity is 0, i.e. the class of numbers congruent to zero modulo m. The inverse of a is $-a$.
5. For each $m > 1$, the classes of residues $\{a : (a, m) = 1\}$ form a group under multiplication modulo p. The identity is 1, and each element has a unique inverse (Proposition 5.9.)

In the abstract notation, it is often convenient to write gh for $g \cdot h$.

Lemma 13.3. *In any group G, the identity element $e \in G$ is unique, and each element g has a unique inverse.*

Proof. Suppose e and e' both satisfy the second axiom of a group G. Then $e = e \cdot e'$, since e' is an identity element, but $e \cdot e' = e'$ since e is identity, therefore necessarily $e = e'$.

Suppose h, h' are inverses to $g \in G$, satisfying $gh = hg = e$ and $gh' = h'g = e$. Then

$$h' = h'e = h'(gh'') = (h'g)h'' = eh'' = h''.$$

□

Every example of a group we gave above satisfies $gh = hg$ for $g, h \in G$. Such groups are called *commutative.* We will not need examples of non-commutative groups.

A group G is called *cyclic*, if it is generated by one element. In other words, if there exists some $g \in G$ such that $G = \{g, g^2, g^3, \cdots\}$. By the primitive root theorem, we know that the only cyclic groups of type (5) in the examples given above are those with $m = 2, 4, p^\alpha$, or $2p^\alpha$.

Note that several of the cryptosystems defined previously simply utilize the fact that $\{1, \cdots, p-1\}$ modulo p form a cyclic group under multiplication. Where we picked a prime p and a primitive root g modulo p, we might have picked instead any finite cyclic group G, with the choice of a generator.

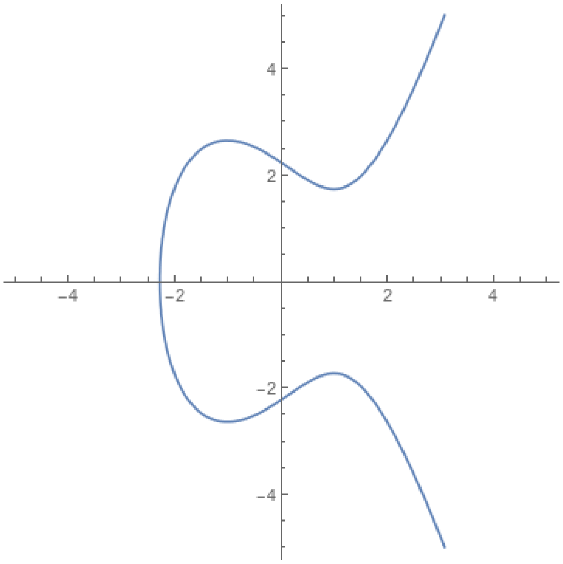
We will now describe an example of such a group, with applications to cryptography.

Elliptic Curves

Let a, b be real numbers that the polynomial $x^3 + ax + b$ has no repeated roots. An *elliptic curve* can be written as an equation of the form

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

If x, y represent real numbers, the set of solutions (x, y) to this equation is indeed a curve in the xy-plane.



Example: the elliptic curve $y^2 = x^3 - 3x + 5$.

On the set E, we define an operation $+ : E \times E \to E$ as follows. If $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ are points on the curve E, then $P + Q$ is the point (x_3, y_3) defined by

$$x_3 = s^2 - x_1 - x_2, \quad y_3 = -s(x_3 - x_1) - y_1, \tag{12}$$

where s is the slope of the line connecting P and Q, i.e.

$$s = \frac{y_2 - y_1}{x_2 - x_1}.$$

Let us for now ignore that this is not well-defined if $x_1 = x_2$. The operation $(P, Q) \mapsto P + Q$, captures the following geometric construction:

(1) Draw the straight line passing through P to Q.
(2) The line will intersect E at a third point, let R be that point.
(3) Draw a vertical line through R, and let it intersect the curve again at R'.
(4) The point R' is $P + Q$.

(...incomplete...)

14. Quadratic Residues

Let $a, m \in \mathbb{Z}$ be such that $(a, m) = 1$.

Definition 14.1. We say a is a *quadratic residue* modulo m, if $x^2 \equiv a \pmod{m}$ has a solution. If no solution exists, we say a is a *quadratic nonresidue* modulo m.

By the Chinese Remainder Theorem and Hensel's Lemma, whether $x^2 \equiv a \pmod{m}$ has a solution or not can be reduced to the case where m is a prime.

Let p be an odd prime. For $a \in \mathbb{Z}$, the **Legendre symbol** is defined to be

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

Proposition 14.2. *$x^2 + ax + b \equiv 0$ (mod p) has a solution if and only if $\left(\frac{a^2-4b}{p}\right) = 0$ or 1.*

Proof. We have

$$x^2 + ax + b \equiv 0 \text{ (mod } p) \iff 4x^2 + 4ax + 4b \equiv 0 \text{ (mod } p) \iff (2x+a)^2 - a^2 + 4b \equiv 0 \text{ (mod } p).$$

Note the first step is justified since 2 is invertible modulo an odd prime. For the same reason we can put $y = 2x + a$, and the equation becomes

$$y^2 \equiv a^2 - 4b \equiv 0 \text{ (mod } p).$$

□

Proposition 14.3. *The number of solutions of $x^2 \equiv a$ (mod p) is $\left(\frac{a}{p}\right) + 1$.*

Proof. We check this in each case:

Case 1 : $p|a$

Then $\left(\frac{a}{p}\right) = 0$, and $x^2 \equiv a$ becomes $x^2 \equiv 0$ (mod p). The unique solution is $x \equiv 0$, and indeed

$$1 = \left(\frac{a}{p}\right) + 1.$$

Case 2 : a is a quadratic residue $\Longrightarrow \left(\frac{a}{p}\right) = 1$

$x^2 \equiv a$ (mod p) has a solution x_0. Then $x_0^2 \equiv a$ (mod p), so the congruence equation becomes

$$x^2 \equiv x_0^2 \text{ (mod } p) \iff (x - x_0)(x + x_0) \equiv 0 \text{ (mod } p).$$

Note that $x_0 \not\equiv -x_0$ otherwise

$$2x_0 \equiv 0 \Longrightarrow x_0 \equiv 0 \Longrightarrow a \equiv x_0^2 \equiv 0 \Longrightarrow p \mid a \Longrightarrow (p, a) \neq 1.$$

So there are two solutions $x \equiv x_0$ and $x \equiv -x_0$ (mod p). We have

$$2 = \left(\frac{a}{p}\right) + 1.$$

Case 3 : a is a quadratic nonresidue $\Longrightarrow \left(\frac{a}{p}\right) = -1$

$x^2 \equiv a$ (mod p) has no solution. Then

$$0 = \left(\frac{a}{p}\right) + 1.$$

□

Theorem 14.4 (Euler's Criterion)**.** *$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}}$ (mod p).*

Proof. If $p \mid a$, $\left(\frac{a}{p}\right) = 0$, $a^{\frac{p-1}{2}} \equiv 0$ (mod p). Now suppose $p \nmid a$.

Let g be a primitive root modulo p. Then $a \equiv g^r$ (mod p) for some r. Writing $x \equiv g^y$ (mod p), then $x^2 \equiv a$ (mod p) has a solution if and only if

$$g^{2y} \equiv g^r \text{ (mod } p) \iff 2y \equiv r \text{ (mod } p-1).$$

Since $2 \mid p - 1$, $2y \equiv r$ (mod $p - 1$) has a solution if and only if $2 \mid r$, so

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{if } r \text{ is even} \\ -1 & \text{if } r \text{ is odd} \end{cases}$$

On the other hand

$$a^{\frac{p-1}{2}} \equiv g^{r\frac{p-1}{2}} = \left(g^{\frac{(p-1)}{2}}\right)^r \equiv (-1)^r \pmod p = \begin{cases} 1 & \text{if } r \text{ is even} \\ -1 & \text{if } r \text{ is odd} \end{cases}$$

□

Proposition 14.5. *Let p be an odd prime, $a, b \in \mathbb{Z}$.*

(1) *If $a \equiv b$ (mod p), then $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$*

(2) *If $p \nmid a$, then $\left(\frac{a^2}{p}\right) = 1$. In particular $\left(\frac{1}{p}\right) = 1$.*

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

Proof. (1) This is obvious since if $a \equiv b \pmod p$, then the equation $x^2 \equiv a \pmod p$ is the same as $x^2 \equiv b \pmod p$.

(2) If $p \nmid a \Longrightarrow x^2 \equiv a^2 \pmod p$ has a solution $x = a$, so $\left(\frac{a}{p}\right) = 1$.

(3) $\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) \equiv a^{\frac{p-1}{2}} b^{\frac{p-1}{2}} = (ab)^{\frac{p-1}{2}} \equiv \left(\frac{ab}{p}\right) \pmod p$. So

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

□

Proposition 14.6.

$$\left(\frac{-1}{p}\right) = \begin{cases} 1 & \textit{if } p \equiv 1 \ (\textit{mod } 4) \\ -1 & \textit{if } p \not\equiv 3 \ (\textit{mod } 4) \end{cases}$$

Proof. Follows immediately from Euler's criterion, since

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

□

Corollary 14.7. *There are infinitely many primes $\equiv 1$ (mod 4)*

Proof. Suppose $\{p_1, \cdots, p_r\}$ are distinct primes congruent to 1 modulo 4. Consider

$$A = 4(p_1 p_2 \cdots p_r)^2 + 1,$$

and let p be a prime factor of A. Then

$$4(p_1 p_2 \cdots p_r)^2 \equiv -1 \pmod p \Longrightarrow \left(\frac{-1}{p}\right) = 1 \Longrightarrow p \equiv 1 \pmod 4.$$

Since $A \equiv -1 \pmod{p_i}$ and $A \equiv 0 \pmod p$, p is not an element of $\{p_1, \cdots, p_r\}$. Therefore we can always enlarge a finite set of primes congruent to 1 modulo 4, so there must be infinitely many of them. □

15. Quadratic Reciprocity

We learned that $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$ for odd primes p. Here is an application.

Example 15.1. Claim: The equation $y^2 = x^3 + 7$ has no solutions.

Assume otherwise, and write

$$y^2 + 1 = x^3 + 8 = (x+2)(x^2 - 2x + 4).$$

If $2 \mid x$, then $y^2 = x^3 + 7 \equiv 7 \pmod 8$, which is impossible, since only $0, 1, 4$ are quadratic residues modulo 8. Then $2 \nmid x$.

Now $x^2 - 2x + 4 \equiv 1 - 2 \equiv 3 \pmod 4 \Longrightarrow \exists p \equiv 3 (\text{mod } 4)$ such that $p \mid x^2 - 2x + 4$. Then $p \mid y^2 + 1$, so $y^2 \equiv -1 \pmod p$. But then $\left(\frac{-1}{p}\right) = 1$, which is impossible since $p \equiv 3 \pmod 4$.

Let p be an odd prime, and $a \in \mathbb{Z}$ coprime to p. For each $i \in \{1, \cdots, \frac{p-1}{2}\}$, let r_i be the remainder of dividing $i \cdot a$ by p:

$$ia = q_i p + r_i, \quad 0 \leq r_i < p.$$

Then $\{r_1, \cdots, r_{\frac{p-1}{2}}\}$ can be partitioned into two sets:

$$S = \{s_1, \cdots, s_n\} = \{r_i : r_i > \frac{p}{2}\}, \quad T = \{t_1, \cdots, t_k\} = \{r_i : r_i < \frac{p}{2}\}.$$

We will use the following fact about S and T twice.

Lemma 15.2. *The set* $\{p - s_1, p - s_2, \cdots, p - s_n, t_1, \cdots, t_k\}$ *is the same as* $\{1, \cdots, \frac{p-1}{2}\}$.

Proof. Note that $r_i \neq 0$ since ia is not divisible by p. Then $p - s_i$ and t_j belong to $\{1, \cdots, \frac{p-1}{2}\}$, and so it's enough to show they are distinct. If $i \neq j$ and $r_i = r_j$, then $(i-j)a$ is divisible by p, which isn't possible since $1 \leq i, j \leq \frac{p-1}{2}$. This shows $t_i \neq r_j$ if $i \neq j$ and $p - s_i \neq p - s_j$. It's then enough to show $p - s_i \neq t_j$, or $t_i + s_j \neq p$. This is again impossible since $t_i, s_j \in \{1, \cdots, \frac{p-1}{2}\}$. □

Lemma 15.3 (Gauss)**.** $\left(\frac{a}{p}\right) = (-1)^{\#S}$.

Example 15.4. $p = 7$, $a = 3$.

$$\{a, 2a, 3a\} = \{3, 6, 9\} \equiv \{2, 3, 6\} \pmod 7.$$

$6 > \frac{7}{2}$, but $2, 3 < \frac{7}{2}$, so $\#S = 1$. Indeed, $\left(\frac{3}{7}\right) = (-1)^{\#S} = -1$.

Proof. Let $n = \#S$. Using Lemma 15.2 we compute $(\frac{p-1}{2})!$ modulo p as

$$\begin{aligned} 1 \cdot 2 \cdots \frac{p-1}{2} &= (p - s_1)(p - s_2) \cdots (p - s_n) t_1 t_2 \cdots t_k \\ &\equiv (-s_1)(-s_2) \cdots (-s_n) t_1 t_2 \cdots t_k \\ &\equiv (-1)^n s_1 s_2 \cdots s_n t_1 t_2 \cdots t_k \equiv (-1)^n r_1 r_2 \cdots r_{\frac{p-1}{2}} \\ &\equiv (-1)^n a \cdot 2a \cdot 3a \cdots \frac{p-1}{2} \cdot a \\ &\equiv (-1)^n a^{\frac{p-1}{2}} \cdot 1 \cdot 2 \cdots \frac{p-1}{2} \pmod p \end{aligned}$$

Therefore $a^{\frac{p-1}{2}} \equiv (-1)^n \pmod p \implies \left(\frac{a}{p}\right) = (-1)^n$. □

Corollary 15.5.

$$\left(\frac{2}{p}\right) = \begin{cases} 1 & \textit{if } p \equiv 1 \textit{ or } 7 \ (\textit{mod } 8) \\ -1 & \textit{if } p \equiv 3 \textit{ or } 5 \ (\textit{mod } 8) \end{cases}$$

Proof. Take $a = 2$. Use the same notation from Lemma 15.2.

$$\{a, 2a, \cdots, \frac{p-1}{2} \cdot a\} = \{2, 4, 6, \cdots, p-3, p-1\}$$

So $r_i > \frac{p}{2} \iff 2i > \frac{p}{2} \iff i > \frac{p}{4}$.
Case (i) $p \equiv 1 \pmod 4$: then $2i > \frac{p}{2} \iff i \geq \frac{p+3}{4}$.
So $n = \frac{p-1}{2} - \frac{p-1}{4} = \frac{p-1}{4}$.
$\implies \left(\frac{2}{p}\right) = (-1)^{\frac{p-1}{4}} = \begin{cases} 1 & \text{if } p \equiv 1 \pmod 8 \\ -1 & \text{if } p \equiv 5 \pmod 8 \end{cases}$
Case (ii) $p \equiv 3 \pmod 4$ then: $2i > \frac{p}{2} \iff i \geq \frac{p+1}{4}$.
So $n = \frac{p-1}{2} - \frac{p-3}{4} = \frac{p+1}{4}$.
$\implies \left(\frac{2}{p}\right) = (-1)^{\frac{p+1}{4}} = \begin{cases} 1 & \text{if } p \equiv 7 \pmod 8 \\ -1 & \text{if } p \equiv 3 \pmod 8 \end{cases}$ □

Theorem 15.6 (Quadratic Reciprocity Law)**.** *Let p, q be distinct odd primes. Then*

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

Using this we can compute $\left(\frac{a}{p}\right)$ for any prime p and $a \in \mathbb{Z}$.

Example 15.7. $\left(\frac{126}{191}\right)$
$126 = 2 \cdot 3^2 \cdot 7.$
$\Longrightarrow \left(\frac{126}{191}\right) = \left(\frac{2}{191}\right)\cancel{\left(\frac{3^2}{191}\right)}\left(\frac{7}{191}\right)$

$\left(\frac{2}{191}\right) = 1$ since $191 \equiv 7 \pmod 8$.
$\left(\frac{7}{191}\right)\left(\frac{191}{7}\right) = (-1)^{\frac{191-1}{2}\cdot\frac{7-1}{2}} = -1$
$\Longrightarrow \left(\frac{7}{191}\right) = -\left(\frac{191}{7}\right)$
$\left(\frac{191}{7}\right) = \left(\frac{2}{7}\right) = 1 \quad (\text{ since } 7 \equiv 8 \pmod 8))$
$\Longrightarrow \left(\frac{126}{191}\right) = -1.$

We recall the notation

$$\lfloor x \rfloor = \text{ greatest integer } \leq x.$$

e.g. $\lfloor 4.6 \rfloor = 4, \quad \lfloor -3.1 \rfloor = -4,$ *etc.*

Remark 15.8. The result of the division algorithm can be written as

$$b = a\lfloor\frac{b}{a}\rfloor + r, \quad 0 \leq r < a.$$

Our proof of quadratic reciprocity will use the following expression for $\left(\frac{p}{q}\right)$.

Lemma 15.9. *Let p, q be odd primes. Then*

$$\left(\frac{p}{q}\right) = (-1)^m, \quad \textit{where } m = \sum_{i=1}^{\frac{p-1}{2}} \lfloor\frac{iq}{p}\rfloor.$$

Proof. For $i = 1, 2, \cdots, \frac{p-1}{2}$, write

$$iq = p\lfloor\frac{iq}{p}\rfloor + r_i. \quad (0 \leq r_i < p)$$

As before, let $\{s_1, s_2, \cdots, s_n\} = \{r_i : r_i > \frac{p}{2}\}$, $\{t_1, \cdots, t_k\} = \{r_i, r_i < \frac{p}{2}\}$.
By Lemma 15.2, $\{p - s_1, \cdots, p - s_n, t_1, \cdots, t_k\} = \{1, 2, \cdots, \frac{p-1}{2}\}$.
Then on the one hand

$$\sum_{i=1}^{\frac{p-1}{2}} iq = q\sum_{i=1}^{\frac{p-1}{2}} i = q\left(\sum_{i=1}^{n}(p - s_n) + \sum_{j=1}^{k} t_j\right) = q(pn - \sum_{i=1}^{n} s_i + \sum_{j=1}^{k} t_j),$$

on the other hand

$$\sum_{i=1}^{\frac{p-1}{2}} iq = \sum_{i=1}^{\frac{p-1}{2}} p\lfloor\frac{iq}{p}\rfloor + r_i = p\sum_{i=1}^{\frac{p-1}{2}}\lfloor\frac{iq}{p}\rfloor + \sum s_i + \sum t_j.$$

Therefore

$$pqn - (q+1)\sum_{i=1}^{n} s_i + (q-1)\sum_{j=1}^{k} t_j = p\sum_{i=1}^{\frac{p-1}{2}}\lfloor\frac{iq}{p}\rfloor.$$

Since $p, q \equiv 1 \pmod 2$ and $q \pm 1 \equiv 0 \pmod 2$,

$$n \equiv \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{iq}{p} \rfloor \pmod 2 \implies \left(\frac{q}{p}\right) = (-1)^n = (-1)^m.$$

□

The following proof is due to Eisenstein:

Proof of Quadratic Reciprocity (Theorem 15.6).
By Lemma 15.9, we have

$$\left(\frac{q}{p}\right) = (-1)^{\sum_{i=1}^{\frac{p-1}{2}}} \lfloor \frac{iq}{p} \rfloor, \quad \left(\frac{p}{q}\right) = (-1)^{\sum_{j=1}^{\frac{q-1}{2}} \lfloor \frac{jp}{q} \rfloor}.$$

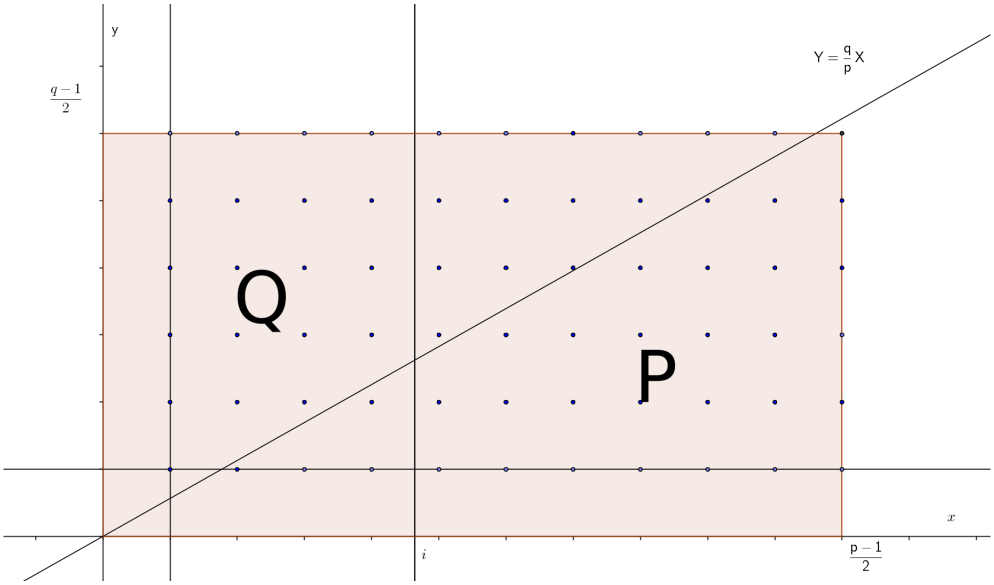
So

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^A, \tag{13}$$

where

$$A = \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{iq}{p} \rfloor + \sum_{j=1}^{\frac{q-1}{2}} \lfloor \frac{jp}{q} \rfloor.$$

Now consider the following picture



Let

$$R = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : 1 \le x \le \frac{p-1}{2}, 1 \le y \le \frac{q-1}{2}\}$$
$$P = R \cap \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : y < \frac{q}{p}x\}$$
$$Q = R \cap \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : y > \frac{q}{p}x\}$$

On the one hand $\#R = \frac{p-1}{2} \cdot \frac{q-1}{2}$, and on the other $\#R = \#P + \#Q$. Now

$$\#P = \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{iq}{p} \rfloor, \quad \#Q = \sum_{i=1}^{\frac{q-1}{2}} \lfloor \frac{jp}{q} \rfloor.$$

Indeed, $\frac{iq}{p}$ is the point on the line $Y = \frac{q}{p}X$ with $x = i$. The number of integer-coordinate points directly below it is therefore $\lfloor \frac{qi}{p} \rfloor$. Similarly, $\lfloor \frac{pj}{q} \rfloor$ is the number of points $(x, y) \in R$ lying above $Y = \frac{q}{p}X$, and having $y = j$.

Therefore

$$A = \#R = \frac{p-1}{2} \cdot \frac{q-1}{2},$$

which together with (13) proves quadratic reciprocity. □

16. Continued Fractions

The idea. Consider how we approximate an unknown quantity with natural numbers. Suppose we have a ruler which we call a *unit*, whose length we take to be 1, and a line segment whose length x we would like to measure. The simplest scenario is if several lengths of the unit fit exactly onto the line, in which case $x = n$ for some $n \in \mathbb{N}$. The next simplest scenario is if b units fit exactly onto a-multiples of the line. In that case we write $x = \frac{a}{b}$.

Now suppose the length x is neither integer nor rational. For the sake of example, let's take x to be secretly equal to π. A crude first approximation to x is the largest multiple of the unit ruler that fits inside the line segment. Experiment will show this is three, so accordingly x is three unit lengths and then some, i.e. $x = 3 + r_0$, with $0 < r_0 < 1$.

To have a better approximation we must measure the error r_0. In order to compare r_0 it with the ruler we might proceed as follows. First we make the segment of length r_0 ten times longer (which can be done geometrically without knowing its exact length.) Then we compare $10r_0$ with the unit. We will find out in our case that $10r_0 = 1 + r_1$, $0 < r_1 < 1$, so that $r_0 = \frac{1}{10} + \frac{r_1}{10}$, and $x = 3 + \frac{1}{10} + \frac{r_1}{10} \approx 3 + \frac{1}{10} = 3.1$. Then to do even better we might multiply r_1 again by 10 and compare it with the unit, and so on. This process leads to the familiar decimal expansion

$$x = 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1000} + \frac{5}{10000} + \cdots = 3.1415...,$$

corresponding to the sequence of approximating fractions

$$3, \ \frac{31}{10}, \ \frac{314}{100}, \ \frac{3141}{1000}, \ \cdots .$$

In fact one common way to *define* a real number such as π is to identify it with such an approximating sequence of rational numbers.

There is a different and still natural method we could have used to measure the error $r_0 = x - 3$ of the first approximation. We could have *inverted* it, producing a segment of length $r_0^{-1} > 1$, and used the ruler to measure that instead. This inversion can also be done geometrically (say with a straight-edge and compass). This way we find that $r_0^{-1} = 7 + r_1$, $0 < r_1 < 1$ so that $x = 3 + \frac{1}{7+r_1} \approx 3 + \frac{1}{7} = \frac{22}{7}$. To do better we can again approximate the error r_1. Drawing a segment of length r_1^{-1} and comparing it with the unit, we find $r_1^{-1} = 15 + r_2$, $0 < r_2 < 1$, so that $x = 3 + \frac{1}{7+\frac{1}{15+r_2}} \approx 3 + \frac{1}{7+\frac{1}{15}} = \frac{333}{106}$. Then we measure r_2^{-1} and so on. This procedure will produce the *continued fraction* expansion

$$x = 3 + \frac{1}{7 + \frac{1}{15+\frac{1}{1+\cdots}}}, \tag{14}$$

corresponding to the sequence of approximating fractions

$$3,\ \frac{22}{7},\ \frac{333}{106},\ \frac{355}{113},\ \frac{103993}{33102},\ \cdots.$$

Continued fractions are in a sense as natural as the decimal digit expansion of a number. They are unwieldy in some ways, for instance it is not straight-forward to add and multiply them. But they are useful in other ways, for example they typically produce better approximating fractions than the decimal expansion.

Finite continued fractions

Definition 16.1. A *finite continued fraction* is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}},$$

for real numbers $a_0, a_1, \cdots, a_n$.

Notation 16.2. Some common notations for the continued fraction above are

$$x = [a_0; a_1, a_2, \cdots a_n], \quad \text{or} \quad x = \langle a_0; a_1, a_2, \cdots a_n\rangle,$$

where the first term a_0 term which occurs outside the main fraction is separated by a semi-colon. Another notation is

$$x = a_0 + \frac{1}{a_1+}\frac{1}{a_2+}\frac{1}{a_3+}\cdots\frac{1}{a_n}.$$

Definition 16.3. A finite continued fraction $\langle a_0; a_1, a_2, \cdots a_n\rangle$ is called *reduced* if $a_n \neq 1$ when $n > 0$. It is called *simple* if $a_k \in \mathbb{N}$ for $k > 0$.

Note that if $a_n = 1$ we have $a_{n-1} + \frac{1}{a_n} = a_n + 1$, so that $\langle a_0; a_1, a_2, \cdots a_n\rangle = \langle a_0; a_1, a_2, \cdots, a_{n-1} + 1\rangle$.

Unless stated otherwise, all continuous fractions of the form $\langle a_0; a_1, \cdots, a_n\rangle$ will be simple and reduced. We also sometimes write $\langle a_0, a_1, \cdots, a_n\rangle$ without the initial semi-colon.

The process of expressing a ratio $\frac{a}{b}$ as a continued fraction is equivalent to the Euclidean algorithm to compute (a, b). For instance, consider $\frac{217}{96}$.

The expression

$$217 = 2 \cdot 96 + 25$$

is equivalent to

$$\frac{217}{96} = 2 + \frac{25}{96} = 2 + \frac{1}{\frac{96}{25}}.$$

Then we are reduced to finding a continued fraction representation for $\frac{96}{25}$. The second step of the Euclidean algorithm gives

$$96 = 3 \cdot 25 + 21$$

from which

$$\frac{96}{25} = 3 + \frac{21}{25} \Longrightarrow \frac{217}{96} = 2 + \frac{1}{3 + \frac{1}{\frac{25}{21}}}.$$

Continuing:

$$\begin{aligned} 25 &= 1 \cdot 21 + 4, \\ 21 &= 5 \cdot 4 + 1, \\ 5 &= 5 \cdot 1 + 0. \end{aligned}$$

We obtain

$$\frac{25}{21} = 1 + \frac{4}{21}, \qquad \frac{21}{4} = 5 + \frac{1}{4}$$

so that

$$\frac{217}{96} = 2 + \frac{1}{3 + \frac{1}{1+\frac{1}{5+\frac{1}{4}}}} = \langle 2; 3, 1, 5, 4\rangle.$$

Proposition 16.4. *Every rational number can be expressed as a simple reduced finite continued fraction.*

Proof. Suppose $r = \frac{a}{b}$, with $a \in \mathbb{Z}$, $b \in \mathbb{N}$. We have $r = \langle q_0; q_1, q_2, \cdots q_n\rangle$, where q_i are the successive quotients in the Euclidean algorithm applied to a and b, starting with $a = q_0 \cdot b + r_0$, $0 \leq r_0 < b$. Since $b > 0$ and $r_0 \geq 0$, the continued fraction is simple.

If at the next-to-last step of the algorithm, the quotient and remainder are equal, i.e. $a - n = q_n \cdot b_n + q_n$, the last step will be $q_n = 1 \cdot q_n + 0$, and the resulting simple continued fraction $\langle q_0; q_1, q_2, \cdots, q_n\rangle$ will not be reduce. Putting $q'_{n-1} = q_{n-1} + 1$, we will obtain simple reduced $\langle q_0; q_1, q_2, \cdots, q'_{n-1}\rangle = \frac{a}{b}$. □

Lemma 16.5. *Suppose that $x = \langle a_0; a_1, \cdots, a_n\rangle$ is simple and reduced. Then $a_0 = \lfloor x \rfloor$.*

Proof. If $n = 0$, then $x = \langle a_0 \rangle = a_0$ and $x = \lfloor x \rfloor$, so the lemma trivially holds.

Now suppose $n > 1$, so that $x = a_0 + \frac{1}{b}$, where $b = \langle a_1, \cdots, a_n\rangle$. If $n = 1$, $b = a_1 > 1$ by the reduced assumption. If $n > 1$, $b = a_1 + \langle a_2, \cdots, a_n\rangle > 1$ since $a_1 \geq 1$, and $\langle a_1, \cdots, a_n\rangle > 0$. In either case $1 < b \Longrightarrow \frac{1}{b} < 1$, so

$$a_0 \leq a_0 + \frac{1}{b} < a_0 + 1 \Longrightarrow \lfloor x \rfloor = a_0.$$

□

Theorem 16.6. *The representation of a rational number as a simple reduced finite continued fraction is unique.*

Proof. Suppose $x = \langle a_0; a_1, a_2, \cdots, a_n\rangle = \langle b_0; b_1, b_2, \cdots, b_m\rangle$, and assume without loss of generality that $n \leq m$. By the lemma we have $a_0 = \lfloor x \rfloor = b_0$. We proceed by induction on n.

<u>$n = 0$:</u> In that case $x = a_0 = b_0$. Then if $m > 0$ we have

$$x = b_0 + \frac{1}{\langle b_1, \cdots, b_m\rangle} = x + \frac{1}{\langle b_1, \cdots, b_m\rangle} \Longrightarrow 0 = \frac{1}{\langle b_1, \cdots, b_m\rangle}$$

which is absurd. Hence $m = 0$, and the continued fractions are $\langle a_0\rangle$ and $\langle b_0\rangle$, which are equal.

Now suppose the theorem holds for $n \leq k$. If $n = k + 1$, we have

$$x = \langle a_0, \cdots, a_{k+1}\rangle = a_0 + \frac{1}{\langle a_1, \cdots, a_{k+1}\rangle} = b_0 + \frac{1}{\langle b_1, \cdots, b_m\rangle}.$$

Then since $a_0 = b_0$, $\langle a_1, \cdots, a_n\rangle = \langle b_1, \cdots, b_m, \rangle$. Then by assumption $m = n$ and $a_i = b_i$ for all By induction the theorem follows. □

We can now speak of *the* continued fraction representation of x, when x is rational. We next proceed to look at truncations of the continued fraction of x, with an eye towards approximating infinite ones.

Definition 16.7. If $x = \langle a_0, \cdots, a_n\rangle$, the mth *convergent* of x, for $m \leq n$, is $\langle a_0, \cdots, a_m\rangle$.

The convergents of any continued fraction $x = \langle a_0, \cdots, a_n\rangle$, not necessarily simple, can be identified as the following sequence of recursively-defined fractions.

Let $(p_{-1}, q_{-1}) = (1, 0)$ and $(p_0, q_0) = (a_0, 1)$. Define a sequence of pairs (p_k, q_k) for $1 \leq k \leq m$ by

$$p_k = a_k \cdot p_{k-1} + p_{k-2}, \quad q_k = a_k \cdot q_{k-1} + q_{k-2}. \tag{15}$$

Proposition 16.8. $\frac{p_m}{q_m} = \langle a_0, \cdots, a_m\rangle$ *for all* $m \leq n$.

Proof. If $m = 0$ we have $\frac{p_m}{q_m} = \frac{a_0}{1} = \langle a_0\rangle$. If $m = 1$, then

$$\frac{p_1}{q_1} = \frac{p_0 \cdot a_1 + p_{-1}}{q_0 \cdot a_1 + q_{-1}} = \frac{a_0 \cdot a_1 + 1}{a_1} = a_0 + \frac{1}{a_1} = \langle a_0, a_1\rangle.$$

Now suppose $m > 1$ and that the proposition holds for smaller values of m. Then

$$\langle a_0, \cdots, a_m\rangle = \langle a_0, \cdots, a_{m-1} + \frac{1}{a_m}\rangle.$$

If we let (p'_k, q'_k) be the analogous sequence for $\langle a_0, \cdots, a_{m-1} + \frac{1}{a_m}\rangle$, then the sequences match for $k < m$, and by the induction assumption

$$\langle a_0, \cdots, a_m\rangle = \frac{p'_{m-1}}{q'_{m-1}} = \frac{(a_{m-1} + a_m^{-1})p_{m-2} + p_{m-3}}{(a_{m-1} + a_m^{-1})q_{m-2} + q_{m-3}} = \frac{a_m(a_{m-1}p_{m-2} + p_{m-3}) + p_{m-2}}{a_m(a_{m-1}q_{m-2} + q_{m-3}) + q_{m-2}}$$
$$= \frac{a_m p_{m-1} + p_{m-2}}{a_m q_{m-1} + q_{m-2}} = \frac{p_m}{q_m}.$$

The proposition then follows by induction on m. □

The convergents $\frac{p_k}{q_k}$ of a simple continued fraction $\langle a_0, \cdots, a_n\rangle$ satisfy remarkable properties. It's clear that p_k, q_k are integers, and $q_k > 0$. In fact, p_k and q_k are always relatively, so that the fraction $\frac{p_k}{q_k}$ is in lowest terms. More specifically, we have:

Proposition 16.9. *Suppose* $\langle a_0, \cdots a_n\rangle$ *is simple, and* (p_k, q_k) *are defined as in (15). Then for all* $0 \leq k \leq n$,

$$p_k q_{k-1} - q_k p_{k-1} = (-1)^{k-1}.$$

Proof. We proceed by induction on k. If $k = 0$, the assertion is $a_0 \cdot 0 - 1 \cdot 1 = (-1)^{-1}$.

Now assume $k > 1$, and $p_{k-1}q_{k-2} - q_{k-1}p_{k-2} = (-1)^{k-2}$. Then

$$p_k q_{k-1} - q_k p_{k-1} = (a_k p_{k-1} + p_{k-2})q_{k-1} - (a_k q_{k-1} + q_{k-2})p_{k-1} = (-1)(p_{k-1}q_{k-2} - q_{k-1}p_{k-2}) = (-1)^{k-1}.$$

The proposition then follows by induction. □

Remark: The above proofs have the advantage of being completely elementary, but perhaps a bit opaque. If we are equipped with some basic linear algebra it's possible to clarify them a bit.

If for $k \geq 0$, we define the matrices

$$P_k = \begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix}, \quad A_k = \begin{pmatrix} a_k & 1 \\ 1 & 0 \end{pmatrix},$$

then the recursive relations in (15) may be written as

$$P_k = P_{k-1} \cdot A_k, \quad P_0 = A_0$$

from which by induction we obtain:

$$P_k = A_0 \cdot A_1 \cdot A_2 \cdots A_k.$$

Now Proposition 16.9 is just the calculation

$$\det P_k = \det(A_0)\det(A_1)\cdots\det(A_k) = (-1)^k.$$

Corollary 16.10. *For each* k, *the integers* p_k, q_k *associated to a simple continued fraction* $\langle a_0, \cdots, a_n\rangle$, *are relatively prime.*

Proof. Proposition 16.9 shows that the equation $p_k \cdot x + q_k \cdot y = 1$ has integer solutions, namely

$$x = (-1)^{k-1} \cdot q_{k-1}, \quad y = (-1)^k \cdot p_{k-1}.$$

□

In particular, knowing the integers p_k and q_k is the same as knowing the fraction $\frac{p_k}{q_k}$. up to sign.

Corollary 16.11. *Let $a, b \in \mathbb{Z}$, $a \neq 0$, $b > 0$, and $d = \gcd(a, b)$. A solution to the equation*

$$ax + by = d,$$

is given by

$$x = (-1)^{m-1} p_{m-1}, \quad y = (-1)^m q_{m-1},$$

where $\frac{p_{m-1}}{q_{m-1}}$ is the next-to-last convergent of $\frac{a}{b} = \langle a_0, \cdots, a_m \rangle$.

Proof. Writing $a = a'd$, $b = b'd$, the equation becomes $a' \cdot x + b' \cdot y = 1$. The last convergent of $\frac{a}{b} = \frac{a'}{b'} = \langle a_0, \cdots, a_m \rangle$ is just $\frac{a'}{b'}$. Then by Proposition 15 we have

$$a' q_{m-1} - b' p_{m-1} = (-1)^{m-1} \Longrightarrow a \cdot (q_{m-1}(-1)^{m-1}) + b \cdot (p_{m-1}(-1)^m) = d.$$

□

17. Infinite Continued Fractions

We would like to make sense of infinite continued fractions, and identities such as

$$\frac{1+\sqrt{3}}{2} = 1 + \cfrac{1}{2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \cdots}}}} = \langle 1, 2, 1, 2, 1, 2, \cdots \rangle.$$

Definition 17.1. Suppose that $\{a_k\}_{k=0}^{\infty}$ is a sequence of real numbers. The infinite continued fraction

$$a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4 + \cdots}}}}$$

also denoted

$$\langle a_0, a_1, a_2, \cdots \rangle$$

is defined to be the limit

$$\lim_{n \to \infty} \langle a_0, \cdots, a_n \rangle,$$

whenever it exists.

For the sake of simplicity we have ignored some pathological cases. To be strictly correct we would have to take care with certain sequences such as $\{a_k\} = \{1, 0, 1, 1, 1, \cdots\}$. In that case $\langle a_0, a_1 \rangle = \langle 1, 0 \rangle$ does not exist, even though the limit does.

Cauchy Sequences

Let us recall some foundational material from analysis. A sequence of real numbers $(a_n) = (a_0, a_1, \cdots)$ is called *Cauchy*, if

$$\lim_{m,n \to \infty} |a_m - a_n| = 0.$$

This means that given any infinitesimal number $\epsilon > 0$, there always exists some large $N > 0$, such that $|a_m - a_n| < \epsilon$ for all $m, n > N$. In other words, the numbers a_n are "close together" for large n.

Let $\mathcal{C}$ be set of all Cauchy sequences $a = \{a_0, a_1, \cdots\}$ of rational numbers. Define an equivalence relation on $\mathcal{C}$ by

$$a \sim b \overset{\text{def}}{\Longleftrightarrow} \lim_{n \to \infty} |a_n - b_n| = 0.$$

Then the real numbers $\mathbb{R}$ may be *defined* as the set of equivalence classes of Cauchy sequences of rational numbers. This means that, for instance, the number $\pi = 3.1415926535 \cdots$ is identified with an equivalence class of Cauchy sequences, a particular element of which is

$$3, 3.1, 3.14, 3.1415, \cdots$$

Operations $+$ and $\cdot$ are defined as follows. Given Cauchy sequences $a = (a_n)$ and $b = (b_n)$, $a + b$ is the sequence $(a_n + b_n)$ and $a \cdot b$ is $(a_n \cdot b_n)$. One has to verify that these are again Cauchy, and that if $a \sim a'$, $b \sim b'$, then $a + b \sim a' + b'$ and $a \cdot b \sim a' \cdot b'$. These verifications are part of standard introductory analysis and we will just take them for granted. The only use we make of all this is the following statement:

A Cauchy sequence (a_n) of rational numbers $a_i \in \mathbb{Q}$, has a limit $a \in \mathbb{R}$.

If $\mathbb{R}$ is identified with equivalence classes of Cauchy sequences, this statement is true *by definition.*

The following proposition says that *simple* infinite continued fractions always exist.

Proposition 17.2. *If* $a_0 \in \mathbb{Z}, a_1, a_2, \cdots \in \mathbb{N}$, *then the limit*

$$\lim_{n\to\infty} \langle a_0, a_1, \cdots, a_n\rangle$$

exists.

Proof. Let $\{p_k, q_k\}$ be the infinite sequence of pairs of integers defined by the recurrence relations (15), so that $\frac{p_k}{q_k} = \langle a_0, \cdots, a_k\rangle$ for all k. By Proposition 16.9 we have

$$p_k q_{k-1} - q_k p_{k-1} = (-1)^{k-1} \Longrightarrow \frac{p_k}{q_k} - \frac{p_{k-1}}{p_{k-1}} = \frac{(-1)^{k-1}}{q_k q_{k-1}}.$$

Then for $m < n$,

$$\begin{aligned}\left|\frac{p_m}{q_m} - \frac{p_n}{q_n}\right| &= \left|\left(\frac{p_m}{q_m} - \frac{p_{m-1}}{q_{m-1}}\right) + \left(\frac{p_{m-1}}{q_{m-1}} - \frac{p_{m-2}}{q_{m-2}}\right) + \cdots + \left(\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right)\right| \\ &\leq \left|\frac{p_m}{q_m} - \frac{p_{m-1}}{q_{m-1}}\right| + \left|\frac{p_{m-1}}{q_{m-1}} - \frac{p_{m-2}}{q_{m-2}}\right| + \cdots + \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| \\ &= \frac{1}{q_m q_{m-1}} + \frac{1}{q_{m-1}q_{m-2}} + \cdots + \frac{1}{q_{n+1}q_n}.\end{aligned}$$

Now we note that the recurrence relation $q_k = a_k q_{k-1} + q_{k-2}$, together with the simplicity assumption $a_k > 0$ imply that $q_k \geq 1$ for all k, and that furthermore $q_{k+1} \geq q_k + 1$. In particular, we have $q_k \geq k$. Then

$$\begin{aligned}\left|\frac{p_m}{q_m} - \frac{p_n}{q_n}\right| &\leq \frac{1}{m(m-1)} + \frac{1}{(m-1)(m-2)} + \cdots + \frac{1}{(n+1)n} \\ &= \left(\frac{1}{m-1} - \frac{1}{m}\right) + \left(\frac{1}{m-2} - \frac{1}{m-1}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n-1}\right) = \frac{1}{n} - \frac{1}{m} < \frac{1}{n}.\end{aligned}$$

Then for $\epsilon > 0$, if $N = \frac{1}{\epsilon}$, then for all $m, n > N$, $|\frac{p_m}{q_m} - \frac{p_n}{q_n}| < \frac{1}{N} = \epsilon$. This shows that the sequence of convergents $(\frac{p_k}{q_k})_{k=0}^{\infty}$ is Cauchy, and hence has a limit. □

Let us attempt to write a continued fraction representation for the real number $\sqrt{3}$. We note that $1 \leq \sqrt{3} \leq 2$, so we start with

$$\sqrt{3} = 1 + (\sqrt{3} - 1),$$

and write

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

Now $1 \leq \frac{1+\sqrt{3}}{2} \leq \frac{3}{2}$, so

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

and so

$$\sqrt{3} = 1 + \frac{1}{\frac{1+\sqrt{3}}{2}} = 1 + \frac{1}{1 + \frac{1}{2+\frac{1}{\frac{1+\sqrt{3}}{2}}}} = 1 + \frac{1}{1 + \frac{1}{2+\frac{1}{1+\frac{1}{2+\frac{1}{\frac{1+\sqrt{3}}{2}}}}}} = \cdots .$$

At this point it is natural to ask whether

$$\sqrt{3} \overset{?}{=} \langle 1, 1, 2, 1, 2, 1, 2, 1, \cdots \rangle,$$

especially since we now know the right-hand side does exist, by Proposition 17.2. We will prove that the equality is indeed true.

Theorem 17.3. *For $x \in \mathbb{R} - \mathbb{Q}$, let a_k, x_k, for $k \geq 0$, be defined by $x_0 = x$, and*

$$a_k = \lfloor x_k \rfloor, \quad x_{k+1} = \frac{1}{\{x_k\}},$$

where $\{y\} = y - \lfloor y \rfloor$. Then $x = \langle a_0, a_1, \cdots \rangle$.

Note that since $x \notin \mathbb{Q}$, the recurrence relation above is well-defined. That is, since x can not be represented as a finite continued fraction, $\{x_k\} = x_k - \lfloor x_k \rfloor$ is never zero, and x_{k+1} always makes sense.

Proof. Since $x_k = a_k + \frac{1}{x_{k+1}}$ we have

$$x = a_0 + \frac{1}{x_1} = a_0 + \frac{1}{a_1 + \frac{1}{x_2}} = a_0 + \frac{1}{a_1 = \frac{1}{a_2 + \frac{1}{x_3}}} = \cdots ,$$

so that (by induction) $x = \langle a_0, a_1, \cdots, a_n, x_{n+1} \rangle$ for all $n \geq 0$. If $\frac{p_m}{q_m}$ are the convergents of the infinite continued fraction $\langle a_0, a_1, \cdots \rangle$, they are also the convergents of $x = \langle a_0, a_1, \cdots, a_n, x_{n+1} \rangle$, for $m \leq n$. As the last convergent is equal to x, we have

$$x = \frac{p_n x_{n+1} + p_{n-1}}{q_n x_{n+1} + q_{n-1}}$$

so that

$$x - \frac{p_n}{q_n} = \frac{p_n x_{n+1} + p_{n-1}}{q_n x_{n+1} + q_{n-1}} - \frac{p_n}{q_n} = \frac{\cancel{q_n p_n x_{n+1}} + q_n p_{n-1} - \cancel{p_n q_n x_{n+1}} - p_n q_{n-1}}{q_n(q_n x_{n+1} + q_{n-1})} = \frac{(-1)^n}{q_n(q_n x_{n+1} + q_{n-1})}$$

by Proposition 16.9.

Now since $x_{n+1} = \frac{1}{x_n - a_n} \geq 0$, we have $q_n x_{n+1} + q_{n-1} \geq q_{n-1} \geq n - 1$, so that

$$0 \leq \left| x - \frac{p_n}{q_n} \right| = \frac{1}{|q_n(q_n x_{n+1} + q_{n-1})|} \leq \frac{1}{n(n-1)}$$

hence $\lim\limits_{n \to \infty} \left| x - \frac{p_n}{q_n} \right| = 0$, as desired. □

Example 17.4. Recalling our initial investigation of π from (16), we can now makes sense of identities such as

$$\pi = 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{1 + \frac{1}{292 + \frac{1}{1 + \cdots}}}}} = \langle 3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, \cdots \rangle. \tag{16}$$

The numbers $3, 7, 15, 1, 292, 1, \cdots$ are the first few values of a_k from Theorem 17.3 in the case $x = \pi$. The first few convergents are

$$3, \ \frac{22}{7}, \ \frac{333}{106}, \ \frac{355}{113}, \ \frac{103993}{33102}, \ \frac{104348}{33215}, \ \cdots .$$

We next want to show that real numbers are approximated by continued fractions in an optimal sense.

Theorem 17.5 (Law of Best Approximations)**.** *Suppose α is a real number with convergent $\frac{p_n}{q_n}$, and $n > 1$. If p, q are integers such that $0 < q < q_n$, and $\frac{p}{q} \neq \frac{p_n}{q_n}$ then*

$$|q_n\alpha - p_n| < |q\alpha - p|.$$

If $\frac{p'}{q'}$ is a reduced fraction, $q \geq q_2$, such that the above inequality is satisfied for (p', q') in place of (p_n, q_n), then $\frac{p'}{q'}$ is a convergent of α.

Proof. □

18. Periodic Continued Fractions

Recall that the continued fraction expansion of $\sqrt{3}$:

$$\sqrt{3} = \langle 1, 1, 2, 1, 2, 1, 2, 1, \cdots, \rangle.$$

Note that the numbers repeat in pairs, but only after the initial term.

Definition 18.1. A continued fraction $\langle a_0, a_1, \cdots \rangle$ is *periodic* if there exists an integer $r > 0$ such that $a_{n+r} = a_n$ for all n sufficiently large.

Every periodic continued fraction then has the form

$$\langle b_0, b_1, \cdots, b_s, a_0, a_1, \cdots, a_{r-1}, a_0, a_1, \cdots, a_{r-1}, a_0, a_1, \cdots, a_{r-1}, \cdots \rangle.$$

Notation: We represent the above by

$$\langle b_0, b_1, \cdots, b_s, \overline{a_0, a_1, \cdots, a_{r-1}} \rangle.$$

Example 18.2.

$$\sqrt{10} = \langle 3, \overline{6} \rangle, \quad 1 + \sqrt{7} = \langle 3, \overline{1, 1, 1, 4} \rangle.$$

Definition 18.3. A *quadratic irrational number* is one of the form

$$\frac{a + b\sqrt{d}}{c}.$$

where a, b, c, d are integers, with $c \neq 0$, d square-free.

Proposition 18.4. *A irrational number α is a quadratic irrational if and only if it's the root of a quadratic polynomial with integer coefficients.*

Proof. If $\alpha = \frac{a+b\sqrt{d}}{c}$, then

$$(c\alpha - a)^2 = b^2 d \Longrightarrow c^2\alpha^2 - 2ac\alpha + a^2 - b^2 d = 0,$$

so that α is a root of $Ax^2 + Bx + C$, where $A = c^2$, $B = -2ac$, $C = a^2 - b^2 d$. Conversely by the quadratic formula, the roots of $Ax^2 + Bx + C$ are quadratic irrationals. □

Note that our examples of periodic continued fractions were all quadratic irrationals. In fact these are the only possible examples.

Theorem 18.5. *An infinite continued fraction $\langle a_0, a_1, \cdots \rangle$ is periodic if and only if it represents a quadratic irrational real number.*

Proof. Suppose

$$y = \langle b_0, b_1, \cdots, b_s, \overline{a_0, a_1, \cdots, a_{r-1}}\rangle.$$

Setting

$$\alpha = \langle \overline{a_0, a_1, \cdots, a_{r-1}}\rangle.$$

we can write

$$\alpha = \langle a_0, a_1, \cdots, a_{r-1}, \alpha\rangle.$$

Then since α is the rth convergent of the right hand side above, so

$$\alpha = \frac{p_r}{q_r} = \frac{\alpha p_{r-1} + p_{r-2}}{\alpha q_{r-1} + q_{r-2}} \Longrightarrow \alpha^2 q_{r-1} + \alpha(q_{r-2} - p_{r-1}) - p_{r-2} = 0,$$

hence α is a quadratic irrational, being the root of

$$P(x) = q_{r-1}x^2 + (q_{r-2} - p_{r-1}) - p_{r-2}.$$

Then

$$y = b_0 + \cfrac{1}{b_1 + \frac{1}{b_2 + \cdots \frac{1}{b_s + \frac{1}{\alpha}}}}$$

can be written as

$$y = \frac{p(\alpha)}{q(\alpha)}$$

with $p(x)$ and $q(x)$ polynomials with integer coefficients. Replacing $p(x)$ and $q(x)$ by their remainders after division by $P(x)$, we can assume they are linear. In other words

$$y = \frac{a\alpha + b}{c\alpha + d}$$

for some integer a, b, c, d, therefore

$$cy\alpha + dy = a\alpha + b \Longrightarrow \alpha(cy - a) = b - dy \Longrightarrow \alpha = \frac{cy - a}{-dy + b}.$$

Now substituting for α in $P(\alpha) = 0$ and cross multiplying, we obtain a quadratic polynomial satisfied by y.

Conversely, suppose that α satisfies

$$a\alpha^2 + b\alpha + c = 0.$$

Let

$$\alpha = \langle a_0, a_1, \cdots\rangle$$

be the continued expansion, and put

$$\beta_n = \langle a_{n+1}, a_{n+2}, \cdots\rangle.$$

Then

$$\alpha = \langle a_0, a_1, \cdots, a_n, \beta_n\rangle$$

for each n, and that

$$\alpha = \frac{\beta_n p_n + p_{n-1}}{\beta_n q_n + q_{n-1}}.$$

Then plugging this into the equation for α we obtain

$$a(\beta_n p_n + p_{n-1})^2 + b(\beta_n p_n + p_{n-1})(\beta_n q_n + q_{n-1}) + c(\beta_n q_n + q_{n-1})^2 = 0$$

from which

$$A_n\beta_n^2 + B_n\beta_n + C_n = 0,$$

where

$$A_n = ap_n^2 + bp_nq_n + cq_n^2,$$
$$B_n = 2ap_np_{n-1} + b(p_nq_{n-1} + p_{n-1}q_n) + 2cq_nq_{n-1}$$

$$C_n = ap_{n-1}^2 + bp_{n-1}q_{n-1} + cq_{n-1}^2 = A_{n-1}.$$

The discriminant of the equation satisfied by β_n is

$$B_n^2 - 4A_nC_n = (2ap_np_{n-1} + b(p_nq_{n-1} + p_{n-1}q_n) + 2cq_nq_{n-1})^2 - 4(ap_n^2 + bp_nq_n + cq_n^2)(ap_{n-1}^2 + bp_{n-1}q_{n-1} + cq_{n-1}^2)$$

this can be simplified to

$$(b^2 - 4ac)(p_nq_{n-1} - q_np_{n-1})^2 = b^2 - 4ac.$$

Now recall that

$$|\alpha - \frac{p_n}{q_n}| < \frac{1}{q_{n-1}q_n}$$

so that

$$|\alpha q_n - p_n| < \frac{1}{q_{n-1}} < \frac{1}{q_n}.$$

Then we can write

$$p_n = \alpha q_n + \frac{\epsilon}{q_n},$$

for $|\epsilon| < 1$. Now

$$A_n = a(\alpha q_n + \frac{\epsilon}{q_n})^2 + b(\alpha q_n + \frac{\epsilon}{q_n})q_n + cq_n^2 = (a\alpha^2 + b\alpha + c)q_n^2 + (2a\alpha + b)\epsilon + a\frac{\epsilon^2}{q_n^2}$$

$$= 2a\alpha\epsilon + b\epsilon + a\frac{\epsilon^2}{q_n^2}$$

which implies

$$|A_n| \le |2a\alpha| + |b| + |a|.$$

This means there are only finitely many possibilities for A_n, and so the same for C_n. Since

$$B_n^2 = 4A_nC_n + (b^2 - 4ac)$$

there are also finitely many possibilities for B_n. Since β_n is a root of $A_nx^2 + B_nx + C_n$, there must exist $r > 0$ such that $\beta_{n+r} = \beta_n$, so that

$$\langle a_{n+1}, a_{n+2}, \cdots \rangle = \langle a_{n+r+1}, a_{n+r+2}, \cdots \rangle,$$

and hence $\langle a_0, a_1, \cdots, \rangle$ is periodic. □

19. Wednesday, May 24

A *quadratic form* is a polynomial function, in several variables, of degree two. We'll consider only the case when there are two variables.

Definition 19.1. A *binary quadratic form* is a function $Q(x, y) = ax^2 + bxy + cy^2$. It is called *integral* if $a, b, c \in \mathbb{Z}$.

For us, a binary quadratic form will always be integral. Let $Q(x, y)$ be such a form. A natural number theoretic question to ask is whether an integer $n \in \mathbb{Z}$ is *represented* by $Q(x, y)$. That means whether the equation

$$Q(x, y) = n$$

has a solution in the integers.

Suppose we set

$$P(x, y) = Q(x + y, y).$$

Then $P(x, y)$ represents n if and only if $Q(x, y)$ does.

Example 19.2. Let

$$Q(x,y) = x^2 + y^2, \; P(x,y) = x^2 + 2xy + 2y^2.$$

Then $P(x,y) = Q(x+y,y)$, and $Q(x,y) = P(x-y,y)$, so solving $Q(x,y) = n$ is equivalent to solving $P(x,y) = n$. For instance, the fact that

$$5 = 1^2 + 2^2 = Q(1,2)$$

is in some sense equivalent to

$$5 = (-1)^2 + 2(-1)2 + 2^2 = P(-1,2).$$

The equivalence above comes down to the fact that the set of numbers (x,y) and $(x+y,y)$ are the *same*, for x, $y \in \mathbb{Z}^2$. If $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$, we can let

$$L = \{x\mathbf{e}_1 + y\mathbf{e}_2 : x, y \in \mathbb{Z}\}.$$

Of course

$$\mathbb{Z}^2 \to L, \; (x,y) \mapsto x\mathbf{e}_1 + y\mathbf{e}_2$$

is a bijection. However, if we set $\mathbf{f_1} = (1,1)$ and $\mathbf{f_2} = (0,1)$, then we also have

$$L = \{x\mathbf{f}_1 + y\mathbf{f}_2 : x, y \in \mathbb{Z}\}$$

so that

$$\mathbb{Z}^2 \to L, \; (x,y) \mapsto x\mathbf{f}_1 + y\mathbf{f}_2$$

is also a bijection. From this point of view, $Q(x,y)$ and $P(x,y) = P(x+y,y)$ are the *same* function $L \to \mathbb{Z}$. Their difference is in the choice of *basis* $\{\mathbf{e}_1, \mathbf{e}_2\}$ or $\{\mathbf{f}_1, \mathbf{f}_2\}$ used to identify L with $\mathbb{Z}^2$.

It therefore makes sense to consider quadratic forms as functions on vectors $\mathbf{v} \in L$, where $L = \{x\mathbf{e}_1 + y\mathbf{e}_2 : x, y \in \mathbb{Z}\}$ for some $\mathbf{e}_1$, $\mathbf{e}_2 \in \mathbb{R}^2$. To erase the choice of a basis entirely, we write $V = \mathbb{R}^2$.

There's a close relation between quadratic forms and symmetric bilinear forms on V.

Definition 19.3. *symmetric bilinear form* $B(x,y)$ is a function satisfying

(1) $B(x,y) = B(y,x)$
(2) $B(ax + x', y) = aB(x,y) + B(x',y)$

Proposition 19.4. *A funcion $Q(x,y)$ is a quadratic form if and only if*

has integer solutions.

20. Reduced Binary Quadratic Forms

Definition 20.1. A binary quadratic form $Q(x,y) = ax^2 + bxy + cy^2$, with discriminant $D \neq \square$, is called **reduced** if:

(1) $-|a| < b \leq |a| \leq |c|$, and
(2) $b \geq 0$ if $|a| = |c|$.

Definition 20.2. Two quadratic forms Q, Q' are said to be **equivalent**, if there exist integers a, b, c, d, with $ad - bc = \pm 1$, such that $Q(ax + by, cx + dy) = Q'(x,y)$. They are called **properly equivalent** if the same is true for $ad - bc = 1$.

Consider

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

applied to M_Q, for $Q = [a,b,c]$. We have

$${}^tRM_QR = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} c & b/2 \\ b/2 & a \end{pmatrix}.$$

so that $Q = [a,b,c]$ is *equivalent* to $[c,b,a]$. We use the notation

$$Q^\iota = [c,b,a]$$

Lemma 20.3. *Let Q, Q' be equivalent binary quadratic forms. Then either Q or Q^ι is properly equivalent to Q'.*

Proof. Suppose Q and Q' are equivalent, but not properly so. Then $M_Q = {}^tPM_{Q'}P$, with $\det(P) = -1$. We have

$$M_{Q^\iota} = {}^tRM_QR = {}^tR^tPM_{Q'}PR = (PR)^tM_{Q'}(PR).$$

Then Q' is properly equivalent to Q^ι, since $\det(PR) = 1$. □

If $Q_1 \sim Q_2$, then $Q_1^\iota \sim Q_2^\iota$. Indeed,

$$Q_1 \sim Q_2 \Longrightarrow M_{Q_2} = {}^tPM_{Q_1}P \Longrightarrow M_{Q_2^\iota} = {}^tR^tPM_{Q_1}PR = {}^tTM_{Q_1^\iota}{}^tT,$$

where $T = R^{-1}PR$.

Then $Q \mapsto Q^\iota$ induces a well-defined map $[Q] \mapsto [Q^\iota]$ on proper equivalence classes of binary quadratic forms of a given discriminant D.

Proposition 20.4. *Let $S = \{[Q_1], \cdots, [Q_n]\}$ be a complete set of proper equivalence classes of binary quadratic forms of discriminant D. Let $S = S_0 \cup T \cup T'$ be a disjoint union, where*

$$[Q] \in S \iff [Q] = [Q^\iota], \quad \textit{and} \quad [Q] \in T \iff [Q^\iota] \in T'.$$

Then $S_0 \cup T$ is a complete set of equivalence classes of binary quadratic forms of discriminant D.

Proposition 20.5. *Every binary quadratic form is equivalent to a reduced form.*

Proof. Let

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

Note $\det(S) = \det(T) = 1$. We also have $T^n = \begin{pmatrix} 1 & -n \\ 0 & 1 \end{pmatrix}$. This can be shown by induction, using:

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & a+b \\ 0 & 1 \end{pmatrix}.$$

Let $Q(x,y) = ax^2 + bxy + cy^2$, with associated matrix

$$M_Q = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}.$$

Then

$${}^tT^nM_QT^n = \begin{pmatrix} 1 & 0 \\ -n & 1 \end{pmatrix}\begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}\begin{pmatrix} 1 & -n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & \frac{b-2an}{2} \\ \frac{b-2an}{2} & an^2 - bn + c \end{pmatrix}.$$

and

$${}^tSM_QS = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -b/2 & -c \\ a & b/2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} c & -b/2 \\ -b/2 & a \end{pmatrix}.$$

Then T and S transform the coefficients (a,b,c) of $Q(x,y)$ by

$$T^n : (a,b,c) \mapsto (a, b^{-2an}, an^2 - bn + c)$$
$$S : (a,b,c) \mapsto (c, -b, a).$$

We may apply the following algorithm:

(1) If $|b| > |a|$, apply T^{-n} with appropriate n to make $|b| \leq |a|$.
($|b| \equiv b' (\text{mod } 2a)$ for a unique b', $-a < |b'| \leq a$. Take $n = \frac{b-b'}{2a}$.)
(2) If $|a| > |c|$, apply S.
(3) If $|b| \leq |a| \leq |c|$ then stop. Else go to (1).

Claim: This process must stop.

In each loop, look at the coefficient of x^2.

At (1), the coefficient is unchangedx. At (2), if $|a| \leq |c|$, then we have $|b| \leq |a| \leq |c|$, so the process stops. If not, after applying S, the coefficient strictly decreases.

so until we get $|b| \leq |a| \leq |c|$, the coefficient of x^2 strictly decreases. Hence, the process must stop.

As a result, we have $-|a| < |b| \leq |a| \leq |c|$. If $|a| = |c|$, then the resulting form is of the form $ax^2 + bxy \pm ay^2$.

If $b < 0$, apply S: $\pm ax^2 - bxy + ay^2$. Then we get a reduced form. □

Example 20.6. Find a reduced form that is equivalent to $7x^2 + 25xy + 23y^2$.

$$25 \equiv 11 \equiv -3 \pmod{14}.$$

Apply T^2: $(7, 25, 23) \mapsto (7, 25 - 2 \cdot 14, 7 \cdot 2^2 - 25 \cdot 2 + 23) = (7, -3, 1)$

Apply S: $7x^2 - 3xy + y^2 \sim x^2 + 3xy + 7y^2$
so $7x^2 + 25xy + 23y^2 \sim x^2 + 3xy + 7y^2$
$T : x^2 + 3xy + 7y^2 \sim x^2 + xy + 5y^2$

Theorem 20.7. *Assume $D \neq \square$. There exist only finitely many equivalence classes of binary quadratic forms of discriminant D.*

Proof. It suffices to show that there are only finitely many reduced forms of discriminant D.

Case 1: a and c have opposite signs.

Then $ac < 0$, $D = b^2 - 4ac = b^2 + 4|a| \cdot |c| \geq 4a^2$. Therefore $|a| \leq \sqrt{D}/2$.

Case 2: a and c have the same signs.

$D = b^2 - 4ac \leq b^2 - 4|a||c| \leq b^2 - 4a^2 \leq -3a^2$. Then $3a^2 \leq -D$, so $|a| \leq \sqrt{|D|/3}$.

In either case $|a| < \sqrt{|D|}$. so there are finitely many choices for a, and hence also for b. As D is fixed, there is at most one c for each choice of a and b. □

Note that $D = b^2 - 4ac$ is always either either 0 or 1 modulo 4.

Definition 20.8. For a non-square integer D, congruent to 0 or 1 mod 4, we define the *class number* $h(D)$ as the number of equivalence classes of integral binary quadratic forms of discriminant D.

Notation: For convenience, we introduce the notation

$$Q = [a, b, c]$$

meaning

$$Q(x, y) = ax^2 + bxy + cy^2.$$

Recall that a form $[a, b, c]$ is called *primitive* if $gcd(a, b, c) = 1$.

Theorem 20.9. *Any positive-definite binary quadratic form is properly equivalent to a unique reduced form.*

Proof. The exact same proof as in Proposition 20.5 shows that every positive-definite binary quadratic form is properly equivalent to a reduced form. Indeed, the matrices S and T in that proof have determinant 1.

It's then enough to show that two reduced positive-definite binary quadratic forms are properly inequivalent. Suppose on the contrary that $Q = [a, b, c]$ is properly equivalent to $Q' = [a', b', c']$, and that Q, Q' are both reduced. Without loss of generality, may take $a \geq a'$.

Since Q, Q' are positive-definite, $D < 0$ and $a, a' > 0$. Then $|b| \leq a \leq c$ and $|b'| \leq a' \leq c'$, since Q, Q' are reduced.

If $M_{Q'} = {}^tPM_QP$, and

$$P = \begin{pmatrix} \xi & \zeta \\ \eta & \theta \end{pmatrix},$$

then

$$M_{Q'} = \begin{pmatrix} Q(\xi,\eta) & a\xi\zeta + c\eta\theta + \frac{b}{2}(\zeta\eta + \xi\theta) \\ a\xi\zeta + c\eta\theta + \frac{b}{2}(\zeta\eta + \xi\theta) & Q(\zeta,\theta) \end{pmatrix} = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}.$$

In particular, $Q(\xi,\eta) = a'$ for some $\xi,\eta \in \mathbb{Z}$. Then

$$a \geq a' = a\xi^2 + b\xi\eta + c\eta^2 \geq a(\xi^2 + \eta^2) + b\xi\eta$$

Since $b \geq -a$,

$$a \geq a(\xi^2 + \eta^2) - a|\xi\eta| = a(|\xi| - |\eta|)^2 + a|\xi\eta| \geq a|\xi\eta|,$$

which implies $\xi\eta \leq 1$, i.e. $(\xi,\eta) = (\pm1,0),(0,\pm1),(\pm1,\pm1)$ or $(\pm1,\mp1)$.

Case 1: Suppose $\xi = \pm1$, $\eta = 0$, so that $a = a'$. Since $\det(P) = \xi\theta - \zeta\eta = \xi\theta$ must equal 1, we have $\theta = \xi = \pm1$. Then $P = \pm T^m$ for some $m = \pm\zeta$, and

$$M_{Q'} = \begin{pmatrix} a & \frac{b}{2} \pm ma \\ \frac{b}{2} \pm ma & am^2 \pm mb + c \end{pmatrix}.$$

Then $b' = b \pm 2ma$. But if $|b| < a = a'$, this is only possible if $m = 0$ and $b = b'$, therefore $c = c'$ and $Q = Q'$.

Case 2: If $\xi = 0$, $\eta = \pm1$, we have $a \geq a' = c$, which implies $a = a' = c$. In particular $a' \leq a$, so by the same argument as above we have $a' = c'$. From the equality of discriminants, $b^2 = D_Q + 4ac = D_{Q'} + 4a'c' = (b')^2$, so $b = \pm b'$. Since $a = c$, $a' = c'$, and Q, Q' are reduced, we have $b, b' > 0$, therefore $Q = Q'$.

Case 3: Suppose $\eta^2 = \xi^2 = 1$. Then $a' = a \pm b + c$ and $a \geq a' \geq a$, so $a = a'$, $b = \pm c$. But $|b| \leq a \leq c$, so $a = c = |b|$. Then again since Q is reduced, we have $b > 0$, hence $a = b = c$, and $Q = [a,a,a]$. Let $Q_0 = [1,1,1]$, and define Q_0' by $M_{Q_0'} = {}^tPM_{Q_0}P$. It follows that $Q_0(x,y) = \frac{1}{a}Q'(x,y) = x^2 + b_0'xy + c_0'y^2$ is a reduced form equivalent to $\frac{1}{a}Q(x,y) = x^2 + xy + y^2$. It's enough to show $Q_0 = Q_0'$, so we can assume $a = b = c = 1 = a'$.

Now the conditions $|b'| < a' < c'$ imply $b' = 0$ or ±1. From $b^2 - 4ac = -3 = (b')^2 - 4c'$, we get $b' \neq 0$, so $b' = \pm1$, hence $c' = 1$, and $Q'(x,y) = x^2 \pm xy + y^2$. But since $c' = a'$, we must have $b' > 0$, so $Q'(x,y) = x^2 + xy + y^2 = Q(x,y)$.

□

Corollary 20.10. *If $D < 0$, $h^+(D)$ is equal to the number of reduced primitive binary quadratic forms of discriminant D.*

We compute some examples of class numbers for negative discriminants.

Let $H^+(D)$ denote the set of positive-definite primitive reduced binary quadratic forms of discrminant D. Then the class number $h(D)$ is $|H^+(D)|$. Let us compute some examples. The following lemma is useful for this purpose.

Lemma 20.11. *If $Q = [a,b,c]$ is positive definite and reduced of discriminant D. Then*

$$3ac \leq |D|.$$

Proof. If $Q = [a,b,c]$ is positive-definite, then $a,c > 0$. If it's reduced,

$$-a < b \leq a \leq c.$$

Then $|b| \leq a$ and $|b| \leq c$ so $b^2 \leq ac$, hence $D + 4ac = b^2 \leq ac$ so $D \leq -3ac$. Then $|D| \geq 3ac$ since $D < 0$. □

The following corollary is immediate from $a \leq c$.

Corollary 20.12. *For Q as in the lemma, $a \leq \sqrt{|D|/3}$.*

The bound above implies that to find reduced positive-definite forms, one only needs to consider a, b, c in $(-\sqrt{|D|/3}, +\sqrt{|D|/3})$, with $a, c > 0$, which can be done with a computer, or by hand for small values of D.

Example 20.13. $\underline{D = -3:}$

From $3a^2 \leq |D| = 3$ we get $a = 1$, hence also $c = 1$ The condition $3ac \leq 3$, immediately implies $a = c = 1$. Then $b^2 - 4 = -3 \Longrightarrow b^2 = 1 \Longrightarrow b = \pm 1$. But $a = c$, so $b > 0$ if $[a, b, c]$ is reduced, therefore $b = 1$. Then the class number $h(-3) = 1$ and

$$H^+(-3) = \{x^2 + xy + y^2\}.$$

Example 20.14. $\underline{D = -4}$

Again $3ac \leq |D| = 4$, together with $a, c > 0$ imply $a = c = 1$. Then $b^2 = 4 - 4 = 0$, and

$$H^+(-4) = \{x^2 + y^2\}, \quad h(-4) = 1.$$

Example 20.15. $\underline{D = -20}$

If $Q \in H^+(-20)$,

$$3ac \leq |D| = 20 \Longrightarrow ac \leq 6.$$

If $a = 1$, then $b^2 = 0$ or 1, so $D = b^2 - 4ac = -4c$ or $1 - 4c$. Since $4 \mid D$, the only possibility is $-20 = -4c \Longrightarrow c = 5$. This gives the form

$$x^2 + 5y^2.$$

If $a = 2$, we get $8c = 4ac = b^2 + 20$. Then $0 \leq b^2 \leq a^2 = 4$ gives $20 \leq 8c \leq 24$, which is only possible if $b^2 = 4$, $8c = 24 \Longrightarrow c = 3$. Since $-a < b \leq a$, $b^2 = 4$, $a = 2$ imply $b = 2$. We obtain a form

$$2x^2 + 2xy + 3y^2.$$

If $a \geq 3$, then $ac \geq 9$ since $c \geq a$, contradicting $ac \leq 6$. Therefore

$$H^+(-20) = \{x^2 + 5y^2,\ 2x^2 + 2xy + 3y^2\},$$

and $h(-20) = 2$.

Example 20.16. $\underline{D = -23}$

For $Q = [a, b, c] \in H^+(-23)$, the bound $3ac \leq |D| = 23$ gives $ac \leq 7$. Since $a^2 \leq ac$, we have $a \leq \sqrt{7}$ so again $a = 1$ or 2.

If $a = 1$, either $b^2 = 0$ or 1, so $-23 = b^2 - 4ac = -4c$ or $1 - 4c$, of which only $1 - 4c = -23$ is possible, with $c = 6$, $b^2 = 1$. Again since here $|b| = |a|$ we must have $b > 0$, so $b = 1$. We get the form

$$x^2 + xy + 6y^2.$$

If $a = 2$, $a^2 \leq ac \leq 7$ implies $2 \leq c \leq \frac{7}{2}$, so $c = 2$ or 3. If $c = 2$, $b^2 = -23 + 16 = -7$ which is absurd. Then $c = 3$, $b^2 = -23 + 4 \cdot 2 \cdot 3 = 1 \Longrightarrow b = \pm 1$. We obtain distinct forms

$$2x^2 + xy + 3y^2,\ \ 2x^2 - xy + 3y^2.$$

Then

$$H^+(-23) = \{x^2 + xy + 6y^2, 2x^2 + xy + 3y^2, 2x^2 - xy + 3y^2\},$$

and $h(-23) = 3$.

Now let's consider the problem of representation by quadratic forms, i.e. diophantine equations of the form

$$Q(x,y) = ax^2 + bxy + cy^2 = n.$$

If Q is positive-definite, then $a > 0$ and the above is equivalent to

$$4a^2x^2+4abxy+4acy^2 = 4an \iff (2ax+by)^2+(4ac-b^2)y^2 = 4an \iff (2ax+by)^2+|D|\cdot y^2 = 4an.$$

It follows that if (x_0, y_0) is a solution, then

$$y_0^2 \leq \frac{4an}{|D|}.$$

Example 20.17. Suppose we want to find if $Q(x,y) = 5$ has a solution for any binary quadratic form of Q discriminant -23.

If so, Q would be positive-definite, and so properly equivalent to one of the three forms from Example 20.16. As properly equivalent forms represent the same integers, we only need to check the three reduced forms from that example.

For $Q(x,y) = x^2 + xy + 6y^2$, we would have $y_0^2 \leq \frac{4an}{|D|} = \frac{20}{23}$, which leaves only $y_0 = 0$. However, $Q_(x_0, 0) = x_0^2 \neq 5$, so there are no solutions for this Q.

For $Q(x,y) = 2x^2 \pm xy + 3y^2$. If $Q(x_0,y_0) = 5$, then $y_0^2 \leq \frac{4an}{|D|} = \frac{40}{24}$, so $y_0 = 0$, or ± 1. Again $y_0 = 0$ has no solutions since $2x_0^2 \neq 5$. If $y_0 = \pm 1$, then $Q(x_0,y_0) - 5 = 2x_0^2 \pm x_0 - 5$. As a quadratic in x_0, it has discriminant 41, hence no rational or integer roots.

It follows that 5 can not be represented by any binary quadratic form of discriminant -23.

We consider the special case of $D = -4$.

Proposition 20.18. *Let p be an odd prime. Then $x^2 + y^2 = p$ has a solution if and only if $p \equiv 1 \ (mod\ 4)$*

Proof. We have seen one direction before. If $p = x^2 + y^2$, then one of x and y must be even, the other odd, so one of x^2, y^2 must be 0 (mod 4), the other 1 (mod 4).

Suppose conversely that $p \equiv 1 \pmod 4$. Then $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}} = 1$, so -1 is a square modulo p. Let m, k be integers such that $k^2 + 1 = mp$. Then the quadratic form

$$Q(x,y) = px^2 + 2kxy + my^2$$

represents p. Indeed $Q(1,0) = p$. On the other hand the discriminant of Q is

$$4k^2 - 4mp = 4(k^2 - mp) = -4.$$

Now the class number $h(-4)$ is 1, so Q must be properly equivalent to the unique element of $H^+(-4)$, namely $Q_0(x,y) = x^2 + y^2$. Since equivalent forms represent the same integers, Q_0 must also represent p. Therefore $p = x^2 + y^2$ for some $x, y \in \mathbb{Z}$. □

We can then finally settle the problem of which numbers are representable as $x^2 + y^2$. Recall from the assignments the algebraic identity

$$(x^2 + y^2)(u^2 + v^2) = (xu + yv)^2 + (xy - yv)^2. \tag{17}$$

An immediate consequence is that a product of numbers representable by $x^2 + y^2$ are also representable.

Lemma 20.19. *If a prime $q \equiv 3 \ (mod\ 4)$ divides $n = a^2+b^2$, then the power of q in the factorization of n must be even.*

Proof. If $a^2 + b^2 \equiv 0 \pmod q$ and $q \nmid b$, then $(ab^{-1})^2 \equiv -1 \pmod q$, which is impossible. So $q \mid b$ and for the same reason $q \mid a$, therefore $q^2 \mid a^2 + b^2$. Now write $a = qa'$, $b = qb'$, $n' = a'^2 + b'^2$. If the power of q in the factorization of n is odd, so is the power of q in the factorization of n'. But proceeding in this way $n, n', n'', etc.$ gives an infinite list of natural numbers descending in size, all divisible by q. That would be a contradiction. □

Theorem 20.20. *A number $n \in \mathbb{N}$ can be written as $x^2 + y^2$ if and only if every prime $q \mid n$ such that $q \equiv 3(mod\ 4)$ occurs in the prime factorization of n with an even power.*

Proof. If $q \equiv 3(\text{mod } 4)$, the even powers of q are representable by $x^2 + y^2$, simply because squares are representable by it. If

$$n = p_1^{\alpha_1} \cdots p_r^{\alpha_r} q_1^{2\beta_1} \cdots q_s^{2\beta_s}$$

where $p_i \equiv 1, 2 \pmod 4$ and $q_j \equiv 3 \pmod 4$, then $p_i^{\alpha_i}$ and $q_j^{2\beta_j}$ are representable by $x^2 + y^2$, hence so are their products, by (17).

By the lemma conversely, every number that's a sum of two squares must have that form. □

Example 20.21. $2016 = 2^5 \cdot 3^2 \cdot 7$. Since the power of 7 is odd, 2016 is *not* representable as a sum of two squares.

On the other hand, $1234 = 2 \cdot 617$, and $617 \equiv 1 \pmod 4$. Therefore $1234 = x^2 + y^2$ must have solutions. Indeed $1234 = 3^2 + 35^2$.

21. Representation by quadratic forms

The proof of the criterion for a number to be a sum of two squares relied on several fundamental ingredients:

1. For an odd prime p, p is a sum of two squares if and only if $p \equiv 1 \pmod 4$. The proof of the "only if" direction is basic modular arithmetic, and the "if" direction relied on the facts:
   (a) If $\left(\frac{-1}{p}\right) = 1$, p is representable by a quadratic form of discriminant -4.
   (b) Every quadratic form of discriminant -4 is properly equivalent to $x^2 + y^2$.
2. The product of sums of squares is a sum of squares, because
$$(x^2 + y^2)(u^2 + v^2) = (xu + yv)^2 + (xv - yu)^2.$$
3. If n is a sum of two squares, and $q \equiv 3 \pmod 4$ is an odd prime dividing n, then the exponent of q in the factorization of n is even.

These four main ideas may be called:

- Reciprocity Law
- Reduction Theory
- Composition of Forms
- Infinite Descent

To various degrees, these ideas all can be generalized, and then combined to give analogues of the theorem about the sum of squares. The form $x^2 + y^2$ is special in some ways, for instance having a discriminant with class number 1, so the general theorems are not as concise, but they nevertheless have the same flavor.

Recall that we say an integer n is *properly represented* by a form $Q(x, y)$ if $n = Q(p, q)$ for relatively prime p and q. The relation between proper representation and proper equivalence is as follows:

Lemma 21.1. *An integer m is properly represented by a form Q if and only if Q is properly equivalent to $[m, b, c]$ for some $b, c \in \mathbb{Z}$.*

Proof. Suppose Q is a form such that $Q(p,q) = m$, with $(p,q) = 1$. Then there exist r, s such that $rp - sq = 1$. In other words

$$P = \begin{pmatrix} q & s \\ p & r \end{pmatrix}$$

has determinant 1. Then the top left entry of $A = {}^tPM_QP$ is $Q(p,q) = m$, so $B = [m,b,c]$ for some $b, c \in \mathbb{Z}$, and Q_A is properly equivalent to Q. Conversely, if $Q = [m,b,c]$ then $Q(1,0) = m$. □

The lemma above, combined with the one below, is the link between residues and representability by quadratic forms.

Lemma 21.2. *Let $D \equiv 0$ or 1 $(mod\, 4)$, and m an odd integer coprime to D. There exists a primitive quadratic form $Q(x,y)$ of discriminant D properly representing m, if and only if D is a quadratic residue modulo m.*

Proof. If m is properly representable by some form Q of discriminant D. By the previous lemma we can assume $Q = [m,b,c]$ so that $D = b^2 - 4mc$, hence $D \equiv b^2 \pmod{m}$.

Conversely, suppose $D \equiv b^2 \pmod{m}$. We can assume $D \equiv b \pmod{2}$, since $b^2 \equiv (b+m)^2 \pmod{m}$, and $b \not\equiv b + m \pmod{2}$ as m is odd. Then D and b^2 are either both 0 or both 1 (mod 4). Then writing $D - b^2 = km$, we must have have $4|km \implies 4|m$ since m is odd. Then $D - b^2 = 4mc$. Then $Q = [m,b,c]$ has discriminant D, and $Q(1,0) = m$. Since $(m,D) = 1$, $(m,b) = 1$, therefore Q properly represents m. □

We have already proven that every primitive quadratic form is properly equivalent to a reduced form, uniquely so if the form is positive definite. This is the general form of what was call the reduction theory step.

Interesting complications arise when we try to generalize the identity

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

For instance, suppose $D = -20$, so that $h(D) = 2$, corresponding to the forms

$$\{2x^2 + 2xy + 3y^2,\ x^2 + 5y^2\}.$$

If m is properly representable by a form of discriminant -20, it must be representable by one of the above forms. We have

$$(x^2 + 5y^2)(u^2 + 5v^2) = x^2u^2 + 5x^2v^2 + 5y^2u^2 + 25y^2v^2 = (xu + 5yv)^2 + 5(xv - yu)^2$$

So for one of the forms we find a suitable generalization. However, for the other form we obtain:

$$(2x^2 + 2xy + 3y^2)(2u^2 + 2uv + 3v^2) = (2xu + xv + yu + 3yv)^2 + 5(xv - yu)^2.$$

So the product of two numbers represented by $2x^2+2xy+3y^2$ is a number represented by x^2+5y^2!

This fact led Gauss to describe a *composition law* for quadratic forms. In fact, he was defining a *group structure* on the set of reduced primitive binary quadratic form of a given discriminant $D < 0$.

We were able to give an exact characterization of integer solutions to

$$x^2 + y^2 = n,$$

for any $n > 0$. This arose from a combination of several key facts: The fact that $h(-4) = 1$ was one key step. Another was the representability of odd primes by

Let $-D$ be a discriminant, for $D > 0$, so that $-D \equiv 0$ or $1 \pmod 4$. Assume that $h(D) = 1$. Then every quadratic form $Q(x, y)$ of discriminant $-D$ is properly equivalent to

$$x^2 + Dy^2.$$

Then $Q(x, y) = n$ has a solution if and only if $x^2 + Dy^2 = n$ does.

Suppose $n = mr^2$, for some $m, r \in \mathbb{N}$. If $x^2 + Dy^2 = m$ has a solution (x_0, y_0), then $x^2 + Dy^2 = n$ has a solution (rx_0, ry_0).

If $p \mid n$, and $x^2 + Dy^2 = n$, then

$$x^2 + Dy^2 \equiv 0 \pmod p.$$

If $x \neq 0$,

22. Pell's Equation

We now consider *Pell's equation:*

$$x^2 - Dy^2 = 1, \tag{18}$$

with D a positive square-free number.

In other words, we are asking if $1 \in \mathbb{N}$ is representable by the indefinite quadratic form $Q(x, y) = x^2 - Dy^2$.

From the general identity

$$(x^2 - Dy^2)(z^2 - Dw^2) = (xz + Dyw)^2 - D(xw + yz)^2 \tag{19}$$

we see that if (x, y), (z, w) are solutions of (18), so is $(xz+Dyw, xw+yz)$. Note that if $u = x+\sqrt{D}y$ and $v = z + \sqrt{D}w$, then

$$u \cdot v = (xz + Dyw) + (xw + yz)\sqrt{D} \tag{20}$$

One method of computing solutions to (18) uses continued fractions. It is based on the following fact.

Proposition 22.1. *The continued fraction of $\sqrt{D}$ is periodic, and has the specific form*

$$\sqrt{D} = \langle a_0; \overline{a_1, a_2, \cdots, a_{n-1}, 2a_0}\rangle$$

Example 22.2. We have

$$\sqrt{2} = \langle 1; \overline{2}\rangle, \quad \sqrt{3} = \langle 1; \overline{1, 2}\rangle, \quad \sqrt{5} = \langle 2; \overline{4}\rangle, \quad \sqrt{7} = \langle 2, \overline{1, 1, 1, 4}\rangle.$$

and

$$\sqrt{347} = \langle 18, \overline{1, 1, 1, 2, 4, 1, 17, 1, 4, 2, 1, 1, 1, 36}\rangle.$$

Let

$$\sqrt{D} = \langle a_0; \overline{a_1, a_2, \cdots, a_{n-1}, 2a_0}\rangle$$

as in the proposition, and put

$$\delta = a_0 + \sqrt{D} = \langle \overline{2a_0, a_1, a_2, \cdots, a_{n-1}}\rangle.$$

Then

$$\sqrt{D} = \langle a_0; a_1, a_2, \cdots, a_{n-1}, \delta\rangle.$$

Let $\frac{p_k}{q_k}$ be the kth convergent of the above.

Proposition 22.3. *Let n be the length of the period of the continued fraction of $\sqrt{D}$, and $\frac{p_n}{q_n}$ its nth convergent. Then*

$$p_n^2 - Dq_n^2 = (-1)^{n-1}.$$

Proof. Since it's a finite continued fraction, we have $\frac{p_{n+1}}{q_{n+1}} = \sqrt{D}$. By the recursive relations for $\{p_k, q_k\}$,

$$\sqrt{D} = \frac{p_{n+1}}{q_{n+1}} = \frac{\delta \cdot p_n + p_{n-1}}{\delta \cdot q_n + q_{n-1}}.$$

Then

$$\sqrt{D} \cdot (\sqrt{D} + a_0) \cdot q_n + \sqrt{D} \cdot q_{n-1} = (\sqrt{D} + a_0) \cdot p_n + p_{n-1}, \tag{21}$$

$$\implies D \cdot q_n + \sqrt{D} \cdot a_0 q_n + \sqrt{D} q_{n-1} = \sqrt{D} \cdot p_n + a_0 \cdot p_n + p_{n-1}, \tag{22}$$

which implies

$$p_{n-1} = D \cdot q_n - a_0 \cdot p_n, \quad q_{n-1} = p_n - a_0 \cdot q_n.$$

Substituting for p_{n-1} and q_{n-1} into the general relation

$$p_n q_{n-1} - q_n p_{n-1} = (-1)^{n-1}$$

we get

$$p_n(p_n - a_0 \cdot q_n) - q_n(D \cdot q_n - a_0 \cdot p_n) = (-1)^{n-1}$$

□

Example 22.4. The first two convergents of $\sqrt{5} = \langle 2, \overline{4} \rangle$ are $\frac{p_1}{q_1} = 2$ and $\frac{p_2}{q_2} = 2 + \frac{1}{4} = \frac{9}{4}$. We have

$$9^2 - 5 \cdot 4^2 = 1.$$

The first few convergents of $\sqrt{13} = \langle 3, \overline{1, 1, 1, 1, 6} \rangle$ are

$$3,\ 4,\ \frac{7}{2},\ \frac{11}{3},\ \frac{18}{5},\ \frac{119}{33}, \cdots$$

and we have

$$18^2 - 13 \cdot 5^2 = -1.$$

If the continued fraction of $\sqrt{D}$ has even period n, Proposition 22.3 shows (p_n, q_n) is a solution to Pell's equation (18). If the period is odd, putting $(x, y) = (z, w) = (p_n, q_n)$ in (19) we obtain

$$(p_n^2 + Dq_n^2)^2 - D \cdot (2p_n q_n)^2 = 1.$$

Example 22.5. We have $\sqrt{41} = \langle 6, \overline{2, 2, 12} \rangle$, with convergents

$$6,\ \frac{13}{2},\ \frac{32}{5},\ \frac{397}{62},\ \frac{826}{129}, \cdots .$$

Then $(p_3, q_3) = (32, 5)$, and we have

$$p_3^2 - 41q_3^2 = 32^2 - 41 \cdot 5^2 = -1,$$

$$(p_3^2 + 41{q_3}^2)^2 - D \cdot (2p_3 q_3)^2 = 2049^2 - 41 \cdot 320^2 = 1.$$

So we obtain $(2049, 320)$ as a solution to Pell's equation with $D = 41$.

Definition 22.6. A *fundamental solution* to $x^2 - Dy^2 = \pm 1$, is a pair of solutions (a, b), with a and b positive integers, and a minimal.

The main theorem relating Pell's equation and continued fractions is the following.

Theorem 22.7. *A fundamental solution for $x^2 - Dy^2 = \pm 1$ always exists, and is equal to (p_n, q_n), where $\frac{p_n}{q_n}$ is the nth convergent of the continued fraction representation of $\sqrt{D}$.*

Since $p_n^2 - Dq_n^2 = (-1)^n$, for the equation $x^2 - Dy^2 = -1$ to have a solution it's sufficient that n be odd. In fact, this is also a necessary condition.

Proposition 22.8. *Suppose that $\frac{p_n}{q_n}$ is the n convergent of $\alpha = \langle a_0; a_1, a_2, \cdots \rangle$. If $p, q \in \mathbb{Z}_+$ satisfy*

$$|p - \alpha q| \leq |p_n - \alpha q_n|, \quad q < q_{n+1},$$

then $p = p_n$, $q = q_n$.

Proof. We give a geometric proof, by considering the distance between the line $y = \alpha x$, and the points (q, p) on the lattice of integers $\mathbb{Z}^2 \subset \mathbb{R}^2$.

In general, the shortest distance between a line $Ax + By + C = 0$ and a point (x_0, y_0) is given by the formula

$$\frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}}.$$

Then the distance between $\alpha x - y = 0$ and (q, p) is

$$\frac{|\alpha q - p|}{\sqrt{\alpha^2 + 1}}.$$

Then up to scaling by the factor $\sqrt{\alpha^2 + 1}$, which doesn't depend on the point (q, p), the distance is $|\alpha q - p|$.

Now recall that $p_{n+1}q_n - q_{n+1}p_n = (-1)^n$, so that the

$$\begin{pmatrix} q_n & q_{n+1} \\ p_n & p_{n+1} \end{pmatrix}$$

has determinant ±1. This implies that (q_n, p_n) and (q_{n+1}, p_{n+1}) form a basis for $\mathbb{Z}^2$. In particular, $(q, p) = r(q_n, p_n) + s(q_{n+1}, p_{n+1})$ for $r, s \in \mathbb{Z}$. Now plugging in

$$q = rq_n + sq_{n+1}, \quad p = rp_n + sp_{n+1}$$

into $|\alpha q - p|$ we get

$$|\alpha q - p| = |\alpha r q_n + \alpha s q_{n+1} - rp_n - sp_{n+1}| = |r(\alpha q_n - p_n) + s(\alpha q_{n+1} - p_{n+1})|.$$

Now since $q < q_{n+1}$, $q = rq_n + sq_{n+1}$ implies $rs \leq 0$. At the same time, we know that $\alpha q_n - p_n$ and $\alpha q_{n+1} - p_{n+1}$ have opposite signs. This implies

$$|\alpha q - p| = |r| \cdot |\alpha q_n - p_n| + |s| \cdot |\alpha q_{n+1} - p_{n+1}|.$$

Since by assumption $|\alpha q - p| \leq |\alpha q_n - p_n|$, we must have either $|r| = 1$, $s = 0$, or $r = 0$. If $r = 0$ then $q = sq_{n+1}$, which contradicts $0 < q < q_{n+1}$. From $s = 0$ we have $q = rq_n$. As $q, q_n > 0$, $r > 0$, so $r = 1$. Therefore $q = q_n$, $p = p_n$. □

Theorem 22.9. *Suppose $\alpha = \langle a_0; a_1, a_2, \cdots \rangle$, and $p, q \in \mathbb{Z}$ satisfy*

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

Then $\frac{p}{q}$ is one of the convergents of α.

Proof. We may assume p, q are relatively prime, with $q > 0$. Let n be such that $q_n \leq q < q_{n+1}$. We may assume $|q\alpha - p| \geq |q_n\alpha - p_n|$ since otherwise $(p, q) = (p_n, q_n)$ by Proposition 22.8. Then

$$|\alpha - \frac{p_n}{q_n}| \leq \frac{q}{q_n}|\alpha - \frac{p}{q}| < \frac{1}{2qq_n}.$$

Now

$$|\frac{p}{q} - \frac{p_n}{q_n}| \leq |\alpha - \frac{p}{q}| + |\alpha - \frac{p_n}{q_n}| < \frac{1}{2q^2} + \frac{1}{2qq_n} < \frac{1}{qq_n},$$

so

$$|\frac{pq_n - qp_n}{qq_n}| < \frac{1}{qq_n} \Longrightarrow |pq_n - qp_n| < 1 \Longrightarrow pq_n - qp_n = 0 \Longrightarrow \frac{p}{q} = \frac{p_n}{q_n}.$$

□

We give one application of this theorem to RSA encryption.

Let p, q be distinct odd primes, $n = pq$, $\lambda = \operatorname{lcm}(p-1, q-1)$, and $g = \gcd(p-1, q-1)$. Suppose e, d satisfy $de \equiv 1 \pmod{\lambda}$, so that (n, e), (n, d) are public and private keys for RSA.

Given (n, e), we know that $de = k\varphi(n) + 1$ for some integer k, hence

$$\frac{e}{\varphi(n)} - \frac{k}{d} = \frac{1}{d\varphi(n)}.$$

Therefore $\frac{k}{d}$ is a good approximation of $\frac{e}{\varphi(n)}$. Now the idea is to approximate $\frac{e}{\varphi(n)}$ with $\frac{e}{n}$. Note that $\varphi(n) = (p-1)(q-1) = n - p - q + 1$, so

$$|\frac{e}{n} - \frac{k}{d}| < |\frac{de - kn}{dn}| = |\frac{de - k\varphi(n) + k\varphi(n) - kn}{dn}| = |\frac{1 + k(\varphi(n) - n)}{nd}| = |\frac{1 - k(p+q-1)}{nd}|.$$

Then

$$|\frac{e}{n} - \frac{k}{d}| < \frac{1}{d} \cdot \frac{k(p+q-1)}{n}.$$

Now if $\frac{k(p+q-1)}{n} < \frac{1}{2d}$, the above, along with the theorem implies that $\frac{k}{d}$ is one of the convergents of $\frac{e}{n}$, which is public information.

The inequality is satisfied for instance, if $q < p < 2q$ and $d < \frac{n^{1/4}}{3}$.

Now the idea is that if $n = pq$ can be used to approximate $\varphi(n) = (p-1)(q-1)$, then $\frac{k}{d}$

Appendix: The principle of mathematical induction

Mathematical induction is a recursive method of proof for proving a general mathematical statement that depends on some integer variable n. A typical example is the identity

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

sometimes called Faulhaber's formula. It can be considered as a sequence of statements P_n, starting with

$$\begin{aligned} P_1 : \quad & 1 = \frac{1(1+1)}{2} \\ P_2 : \quad & 1 + 2 = \frac{2(2+1)}{2} \\ P_3 : \quad & 1 + 2 + 3 = \frac{3(3+1)}{2} \\ & \vdots \end{aligned}$$

The idea of induction is to prove P_n for all n by proving the following two statements:

(1) *Base case:* P_1 is true .

(2) *Induction step:* *If* P_k is true for some k, *then* P_{k+1} is also true.

The *principle of mathematical induction* says that, given a sequence of statements P_1, P_2, P_3, $\cdots$, in order to prove P_n is true for all $n \geq 1$, it's enough to prove the two statements above. The first one is usually easy, so the heart of a proof by induction is the second step.

Let us prove Faulhaber's formula by induction, as an example.

Base case: Since $\frac{1(1+1)}{2} = 1$, P_1 is true. That's all there is to the base case.

Induction step: Assume P_k is true, so that

$$1+2+\cdots+k=\frac{k(k+1)}{2}.$$

Then by using this, we have

$$1+2+\cdots+k+(k+1)=\frac{k(k+1)}{2}+k+1=\frac{k(k+1)+2(k+1)}{2}=\frac{(k+1)(k+2)}{2}.$$

This shows that P_{k+1} must be true. Therefore by induction, P_n is true for all $n \geq 1$.

Sometimes a proof requires use of a slightly stronger assumption in the induction step.

(Strong) Induction Step: If P_1, P_2, $\cdots$, P_k are true, so is P_{k+1}.

Occasionally such proofs are said to use *strong* induction. In fact, the two methods of proof are equivalent. To see this, let Q_n denote the statement:

$$P_1, \cdots, P_n \textit{ are true.}$$

The strong induction step for $\{P_n\}$ is the same as ordinary induction for $\{Q_n\}$. The base case is identical for both sequences, and the truth of P_n for all n is the same as the truth of Q_n for all n.

The following principle is often a direct consequence of the axioms. We assume its truth.

Well-Ordering Principle

The natural numbers are well-ordered: every non-empty subset of $\mathbb{N}$ *has a smallest element.*

Theorem. *The principle of mathematical induction is equivalent to the well-ordering principle.*

Proof. Let us prove the well-ordering principle by induction. Let $S \subset \mathbb{N}$ be a set, and let P_n denote the statement:

The number n *does not belong to* S.

Suppose that S does *not* have a smallest element. We prove that it must be empty, i.e. that P_n is true for all $n \in \mathbb{N}$.

Base case: Since 1 is the smallest element of $\mathbb{N}$, it can not belong to S, as it would also be the smallest in S. Therefore P_1 is true.

Induction step: Assume that $P_1, P_2, \cdots, P_k$ are all true, so that $1, 2, \cdots, k \notin S$. If $k+1 \in S$, it would be the smallest in S, since all the smaller natural numbers are not in S. As S has no smallest element, $k+1 \notin S$, and P_{k+1} is true.

Then by (strong) induction, P_n is true for all n. In particular, $n \notin S$ for any $n \in \mathbb{N}$, and yet $S \subset \mathbb{N}$, so $S = \emptyset$.

Now let us prove the principle of mathematical induction itself using the well-ordering principle. Let $P_1, P_2, \cdots$ be any sequence of statements and suppose: (1) P_1 is true. (2) P_k implies P_{k+1}. Let $S \subset \mathbb{N}$ be the set of natural numbers n for which P_n is *false*. We assume it is non-empty, and derive a contradiction. By the well-ordering principle S must have a smallest element n_0. Since P_1 holds, $n_0 > 1$, so that $n_0 - 1$ is again a natural number. As $n_0 - 1 < n_0$ and n_0 is the smallest in S, $n_0 - 1 \notin S$, therefore $P_{n_0 - 1}$ must be true. But the induction step implies P_{n_0} must also be true, contradicting $n_0 \in S$. Therefore S must be empty, and so P_n is true for all n. □

Finally, we remark that mathematical induction is despite its name purely deductive and not an instance of "inductive reasoning" in the philosophical sense.