M3P14 NUMBER THEORY

TOBY GEE

Contents

0. Introduction

These notes are based on notes by David Helm.

Contact details - toby.gee@imperial.ac.uk.

Office hours are Monday and Thursday 1400-1500 (i.e. before the lecture), in 666.

Homework: there will be exercise sheets on blackboard. If you want to get feedback you can hand work in and I'll mark it. Roughly one class in six will be a problems class, going through solutions to these exercises and answering questions.

Books: not basing this on any particular book, although I will probably look at Baker's "A Concise Introduction to the Theory of Numbers" from time to time.

Hopefully I will post these notes on Blackboard (with a delay).

0.1. **What is this course about?** Roughly speaking number theory is the study of the integers; more specifically, problems in number theory often have a lot to do with primes and divisibility, and include problems about the rational numbers (for example, solving equations in integers or in the rationals).

We will be looking at problems that can be tackled by elementary means - but this doesn't mean easy! Also the statements of problems can be elementary without the solution being elementary, e.g. Fermat's Last Theorem, or even known, e.g. the twin prime conjecture.

We'll start the course with a look at prime numbers and factorisation, and we'll return to primes at the end, too. Typical questions here: how do you tell if a number is prime? How many primes are there $\equiv a \pmod{b}$ for given a, b? How many primes are there less than n? (Gauss: roughly $n/\log n$. This is the prime number theorem, which isn't really elementary, and isn't in this course.)

0.2. **Diophantine equations.** These are polynomial equations, perhaps in several variables, with integer coefficients. We'll see an example later in the lecture. Typical questions:

- Do they have solutions in integers? Rational numbers?
- If yes, are there infinitely many?
- If not, how many?

For example, we'll solve linear equations $nx+my = d$ in the integers - either no solutions or infinitely many. We'll also do some higher degree equations, in particular Pell's equation

$$x^2 - Dy^2 = \pm 1,\ D > 0$$

Note that some clue as to the difficulty of these problems is Matiyasevich's theorem, which says that there is no algorithm to determine whether or not a Diophantine equation has a solution.

0.3. **Modular arithmetic.** Rather than solving equations in the integers you can solve them modulo integers (especially primes). This is only a finite amount of work, and it also tells you a lot about the solutions in the integers.

One example we'll think a lot about is the case of squares. For example, we can think about squares mod 10, i.e. the last digits of squares in base 10: 0, 1, 4, 9, 6, 5, 6, 9, 4, 1, 0, 1, 4, 9...

First useful examples are mod 3 and mod 4. For example, all squares are 0 or 1 (mod 4); indeed $(2n)^2 = 4n^2$ and $(2n+1)^2 = 4(n^2+n)+1$. For example we can easily see that we can't solve $x^2 + y^2 = 2015$ in integers, because already there are no solutions mod 4.

0.4. **Example: Pythagorean triples.** Note: this example is going to use stuff from the first year, and maybe a bit more. Everything in it will be justified again later in the course, so don't worry if you don't follow it.

We want to find all integer solutions to the equation $x^2+y^2 = z^2$; provably you all know $(3, 4, 5)$, maybe $(5, 12, 13)$, and so on. (N.b. much harder is Fermat's Last Theorem: no solutions to $x^n + y^n = z^n$ if $n > 2$.)

Firstly we can assume that (x, y, z) are pairwise coprime, as if a prime divides two of x, y, z it divides the third. Also if x, y are both odd then $x^2 + y^2 \equiv 2 \pmod 4$, which isn't a square, so precisely one is odd (at least one is, as they are coprime), say x odd and y even (so z is also odd).

Then rearrange as $y^2 = (z + x)(z - x)$ or

$$(y/2)^2 = (z + x)/2 \cdot (z - x)/2.$$

The two terms on the right hand side are coprime (as a common factor would divide their sum and difference), so both are squares (think about their prime factorisations). So we can write $z + x = 2a^2$, $z - x = 2b^2$, $y = 2ab$ where $a, > b \geq 1$ are coprime and of opposite parity. So we get

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

These are all of them (up to swapping x and y, and up to multiplying all of them by positive integers). e.g. we get $(15, 8, 17)$ from $a = 4, b = 1$, and so on.

1. Euclid's algorithm and Unique Factorization

1.1. **Divisibility.**

Definition 1.1. Let a, b be integers. We say that a *divides* b (written $a \mid b$) if there exists an integer n such that $b = na$.

Note that if a, b, c are integers such that $a|b$ and $a|c$, then $a \mid nb + mc$ for any integers n, m (give proof).

Definition 1.2. Let a, b be integers, not both zero. The *greatest common divisor* or *highest common factor* of a and b (written (a, b)) is the largest positive integer dividing both a and b.

Such an integer always exists since if a is nonzero and $c \mid a$, then $-a \leq c \leq a$.

Definition 1.3. An integer a is *prime* if a has exactly two positive divisors (namely, 1 and a).

Note that by definition, primes can be both positive and negative. In spite of this, in this course when we say "let p be a prime" we will generally mean p is positive.

We will see later in the course that both of these definitions are somewhat naive, and find better formulations of these concepts (i.e. ones that work in much greater generality.)

1.2. **Euclid's algorithm.**

Proposition 1.4. *Let a and b be integers, not both zero. Then for any integer n, we have $(a, b) = (b, a - nb)$.*

Proof. From the definition of (a, b) it suffices to show that any positive integer divides m both a and b if, and only if, it divides both b and $a - nb$. But if m divides a and b, it clearly divides $a - nb$, and if it divides b and $a - nb$, it clearly divides a. □

This suggests an approach to computing (a, b): replace (a, b) by a pair $(b, a-nb)$ that is "smaller", and repeat until the numbers involved are small enough that it is easy to compute the greatest common divisor. The key to being able to do this is the following innocuous looking result:

Theorem 1.5. *Let a and b be integers with $b > 0$. Then there exist integers q and r such that $a = qb + r$ and $0 \leq r < b$.*

Proof. Let $q = \lfloor a/b \rfloor$ be the largest integer less than $\frac{a}{b}$. Then $0 \leq \frac{a}{b} - q < 1$. Thus $0 \leq a - bq < b$, so we can take $r = a - bq$. □

This gives us an algorithm (Euclid's algorithm) for finding (a, b) for any (a, b) not both zero. Without loss of generality (it's fine to change the signs of a, b), assume $0 < b \leq a$ and a is positive.

(1) Check if $b = 0$. If so then $(a, b) = a$.
(2) Otherwise, replace (a, b) with (b, r), where $a = bq + r$ and $0 \leq r < b$, and return to step 1.

Since at every stage $|a|$ is decreasing, this algorithm terminates; we've shown that $(a, b) = (b, r)$ so the output is always equal to (a, b).

Let's make this explicit: $(120, 87) = (87, 33) = (33, 21) = (21, 12) = (12, 9) = (9, 3) = (3, 0) = 3$. Now run this backwards, writing out the equations, to get $3 = 8 \cdot 120 - 11 \cdot 87$.

The same works in general, i.e. the algorithm gives us more than just a way to compute (a, b): it also allows us to express (a, b) in terms of a and b:

Theorem 1.6. *Let a and b be integers, not both zero. Then there exist integers n, m such that $(a, b) = na + mb$.*

Proof. Let $a_0 = a$, $b_0 = b$, and for each i let a_i, b_i be the result after running i steps of Euclid's algorithm on the pair (a, b). For some r we have $a_r = (a, b)$ and $b_r = 0$. We will show, by downwards induction on i, that there exist integers n_i, m_i such that $(a, b) = n_i a_i + m_i b_i$. For $i = r$ this is clear. On the other hand, for any i we have $a_i = b_{i-1}$ and $b_i = a_{i-1} - q_i b_{i-1}$ for some integer q_i. Thus if $(a, b) = n_i a_i + m_i b_i$, we have

$$(a, b) = n_i b_{i-1} + m_i(a_{i-1} - q_i b_{i-1}) = (n_i - m_i q_i)b_{i-1} + m_i a_{i-1}$$

and the claim follows. □

1.3. **Unique factorization.** The fact that (a, b) is an integer linear combination of a and b has strong consequences for factorization and divisibility. First note:

Proposition 1.7. *Let n, a, b be integers, and suppose that $n \mid ab$ and $(n, a) = 1$. Then $n \mid b$.*

Proof. There exist integers x, y such that $ax + ny = 1$. Thus $abx + bny = b$. But n clearly divides abx and bny, so n divides b. □

The following is also useful.

Proposition 1.8. *If $(a, b) = 1$ and n is divisible by both a and b, then it is divisible by ab.*

Proof. Write $n = n(a, b) = n(ax + by) = anx + nby$; each term is divisible by ab. □

By definition, if n is prime, then either n divides a or $(n, a) = 1$. Thus:

Corollary 1.9. *If p is prime, and a, b are integers such that $p \mid ab$, then either $p \mid a$ or $p \mid b$.*

We can now prove the existence and uniqueness of prime factorisations. First existence:

Proposition 1.10. *Every nonzero integer can be written as $\pm p_1 \cdots p_r$ for some $r \geq 0$ and some primes $p_1, \ldots, p_r$.*

Proof. WLOG positive. Then do induction on n: if n isn't prime then $n = ab$ with $1 < a, b < n$ and by induction a, b are both products of primes. □

We can also deduce that factorizations into primes are essentially unique:

Theorem 1.11. *Let n be a positive integer, and suppose we have two factorizations:*

$$n = p_1 p_2 \ldots p_r = q_1 q_2 \ldots q_s,$$

with the p_i and q_j prime (and positive). Then $r = s$ and the p_i are a rearrangement of the q_j.

Proof. Suppose otherwise, and let n be the smallest positive integer where this does not hold. We then have $p_1 \mid q_1 q_2 \ldots q_s$, so either $p_1 \mid q_1$ or $p_1 \mid q_2 q_3 \ldots q_s$. Proceeding inductively, we have $p_1 \mid q_i$ for some i; since q_i is prime this means $p_1 = q_i$. We then have:

$$p_2 p_3 \ldots p_r = q_1 q_2 \ldots q_{i-1} q_{i+1} \ldots q_s.$$

Since this product is smaller than n, we must have that $r - 1 = s - 1$ and the p's (except p_1) are a rearrangement of the q's (except q_i). □

Put together, these are the fundamental theorem of arithmetic.

1.4. **Linear diophantine equations.** Suppose now that we are given a, b, c (all integers) and we want to solve $ax + by = c$ for x, y integers. We first note that (a, b) divides both a and b, so for there to be any solutions, we must have $(a, b) \mid c$. From now on, suppose this is true.

We can write $c = d(a, b)$. We can also find integers m, n such that $(a, b) = ma + nb$. Then $c = d(a, b) = dma + dnb$, so $x = dm$, $y = dn$ is a solution.

Now let x, y be one solution to the equation, and suppose we have a second solution x', y'. Then we find that $a(x - x') + b(y - y') = 0$. We thus have $x' = x + r$, $y' = y + s$, where $ar + bs = 0$. Conversely, if $ar + bs = 0$, then $x + r, y + s$ is a solution to $ax + by = c$. We are thus reduced to finding the solutions to $ar + bs = 0$. For any integer r, to have an s such that $ar + bs = 0$ we must have $s = -\frac{ar}{b}$. So we must determine for which integers r we have $-\frac{ar}{b}$ an integer. This is precisely when $b \mid ar$. Dividing both sides by (a, b), we see that this happens if, and only if, $\frac{b}{(a,b)} \mid \frac{ar}{(a,b)}$. But since $(\frac{b}{(a,b)}, \frac{a}{(a,b)}) = 1$, this happens if, and only if $\frac{b}{(a,b)} \mid r$.

Putting this all together, we find that if x, y is one solution to $ax + by = c$, then the other solutions are of the form $x + n\frac{b}{(a,b)}, y - n\frac{a}{(a,b)}$ for all integers n.

For example, using the example above where we have $8 \cdot 120 - 11 \cdot 87 = 3$, we can solve $120x + 87y = 9$. The solutions are $x = 24 - 29k$, $y = -33 + 40k$; taking $k = 1$, we have for example $x = -5$, $y = 7$.

2. Congruences and Modular Arithmetic

2.1. Congruences.

Definition 2.1. Let n be a nonzero integer (usually taken to be positive) and let a and b be integers. We say a is congruent to b modulo n (written $a \equiv b \pmod{n}$) if $n \mid (a - b)$.

For n fixed, it is easy to verify that congruence mod n is an equivalence relation, and therefore partitions $\mathbb{Z}$ into equivalence classes. We let $[a]_n$ denote the equivalence class of a mod n; that is, the set of b such that $b \equiv a \pmod{n}$. The set of equivalence classes modulo n is denoted $\mathbb{Z}/n\mathbb{Z}$.

For any integer a, we can write $a = qn + r$ with $0 \leq r < n$. Then $a \equiv r \pmod{n}$. It follows that the n congruence classes $[0]_n, \dots, [n-1]_n$ exhaust $\mathbb{Z}$. On the other hand, if $0 \neq r, r' < n$, and $r \equiv r' \pmod{n}$, then $|r - r'| < n$ and $n \mid r - r'$, so $r = r'$. We thus have $\mathbb{Z}/n\mathbb{Z} = \{[0]_n, \dots, [n-1]_n\}$.

It is easy to check that if a, b, c are integers and $a \equiv b \pmod{n}$, then $a + c \equiv b + c \pmod{n}$ and $ac \equiv bc \pmod{n}$. It follows that one can define addition and multiplication on $\mathbb{Z}/n\mathbb{Z}$ by setting:

$$[a]_n + [b]_n = [a + b]_n$$

$$[a]_n[b]_n = [ab]_n.$$

One easily checks that this satisfies the axioms of a ring, and moreover that the map $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ taking an integer a to $[a]_n$ is a ring homomorphism.

Recall that if R is a ring, the *units* of R (denoted $R^\times$) are the elements of R with multiplicative inverses. Let a be an integer and suppose that $[a]_n$ is a unit. Then there exists b such that $ab \equiv 1 \pmod{n}$; it follows that $(a,n) = 1$. Conversely, if $(a,n) = 1$, there exist integers x, y such that $ax + ny = 1$; then $[a]_n[x]_n = [1]_n$, so $[a]_n$ is a unit. Thus we have $(\mathbb{Z}/n\mathbb{Z})^\times = \{[a]_n : (a,n) = 1\}$.

Note that if p is a prime, then either $a \equiv 0 \pmod{p}$ or $(a,p) = 1$. Thus every nonzero congruence class mod p is a unit; i.e. $\mathbb{Z}/p\mathbb{Z}$ is a field.

2.2. **Linear congruence equations.** Fix integers a and c and a nonzero integer n. Suppose we want to solve $ax \equiv b \pmod{n}$. This is equivalent to finding x, y such that $ax + ny = b$. In particular, by our analysis of linear diophantine equations, there is a solution precisely when (a,n) divides b. Moreover, if x is a solution, the other solutions are of the form $x + r\frac{n}{(a,n)}$ for integers r. In particular they form a single congruence class mod $\frac{n}{(a,n)}$, or a total of (a,n) congruence classes mod n.

Example sheet 1 is online now.

2.3. **The Chinese remainder theorem.** Let m and n be nonzero integers. Note that if $a \equiv b \pmod{mn}$, then $a \equiv b \pmod{n}$ and $a \equiv b \pmod{m}$. Thus a congruence class mod mn determines a pair, consisting of a congruence class mod m and a congruence class mod n. More formally, we have a map: $\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ that takes $[a]_{mn}$ to the pair $([a]_m, [a]_n)$. It's easy to check that this map is a ring homomorphism.

A natural question to ask is whether the converse is true; that is, given a pair of congruence classes mod m and mod n, is there a congruence class mod mn giving rise to it? In general the answer will be no, but when $(m,n) = 1$, then we have the so-called Chinese Remainder Theorem:

Theorem 2.2. *Let m and n be integers with $(m,n) = 1$, and let a and b be integers. Then there exists an integer x such that $x \equiv a \pmod{m}$ and $x \equiv b \pmod{n}$. Moreover x is unique modulo mn; that is, if x' is another integer with this property, then $x' \equiv x \pmod{mn}$.*

Proof. We first prove uniqueness. Note that if x and x' both have the required properties, then $x \equiv x' \pmod{n}$ and $x \equiv x' \pmod{m}$. Thus $x - x'$ is divisible by both m and n. Write $x - x' = cm$ for some integers c. We have n divides cm; since $(n,m) = 1$, we must have $n \mid c$. Writing $c = dn$, we have $x - x' = mnd$, so $x \equiv x' \pmod{mn}$.

For existence, we give two proofs. The first is constructive: write $1 = nx + my$. Then $anx + bmy$ is congruent to a mod m and b mod n.

For the second proof, note that the uniqueness we proved above shows that the map:

$$\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$$

is injective. On the other hand both the source and the target of this map have cardinality mn, so the map is surjective as well, which proves the claim. □

Another way to interpret the Chinese remainder theorem is that the map

$$\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$$

is an *isomorphism* of rings. Put another way, to prove any statement about $\mathbb{Z}/mn\mathbb{Z}$ that can be formulated purely in ring-theoretic terms, it suffices to prove corresponding statements mod m and mod n, provided that $(m, n) = 1$.

3. Euler's theorem

3.1. **The Euler Φ-function.** We define a function Φ on positive integers n by letting $\Phi(n)$ denote the number of elements of $(\mathbb{Z}/n\mathbb{Z})^\times$. Explicitly we have $\Phi(n) = \#\{a : 0 \leq a < n, (a, n) = 1\}$; that is, $\Phi(n)$ is the number of integers between 0 and $n - 1$ whose greatest common divisor with n is 1.

The Chinese Remainder theorem shows that if m and n are positive integers with $(m, n) = 1$, we have an isomorphism of rings: $\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, and hence an isomorphism of groups:

$$(\mathbb{Z}/mn\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times.$$

It follows that if $(m, n) = 1$, then $\Phi(mn) = \Phi(m)\Phi(n)$.

Definition 3.1. A function f on the positive integers is *multiplicative* if for all positive integers m, n such that $(m, n) = 1$, we have $f(mn) = f(m)f(n)$. We say f is *strongly multiplicative* if for any pair of positive integers m, n we have $f(mn) = f(m)f(n)$.

Note that Φ is multiplicative but not strongly multiplicative, since for instance $\Phi(2) = 1$ but $\Phi(4) = 2$.

It is clear that a multiplicative function is determined by its values on prime powers. For p prime we have $(a, p^r) = 1$ if, and only if, p does not divide a, so $\Phi(p^r)$ is the number of integers between 0 and $p^r - 1$ that are not divisible by p. There are p^{r-1} numbers in this range divisible by p, so we have $\Phi(p^r) = p^r - p^{r-1} = p^r(1 - \frac{1}{p})$.

From this and multiplicativity of Φ one has that $\Phi(n) = n \prod_{p|n}(1 - \frac{1}{p})$, where p runs over the primes dividing n.

3.2. **Euler's theorem.** Note that the units $(\mathbb{Z}/n\mathbb{Z})^\times$ form a group under multiplication; by definition, $\Phi(n)$ is the order of this group. Recall that for any group G of finite order d, Lagrange's theorem states that for all $g \in G$, g^d is the identity in G. For the group $(\mathbb{Z}/n\mathbb{Z})^\times$, this means that if $[a]_n \in (\mathbb{Z}/n\mathbb{Z})^\times$, we have $[a]_n^{\Phi(n)} = [1]_n$. In other words:

Theorem 3.2. *Let a be an integer with $(a, n) = 1$. Then $a^{\Phi(n)} \equiv 1$ (mod n).*

This is known as *Euler's theorem*. When n is a prime, it becomes the statement that $a^{p-1} \equiv 1 \pmod{p}$ if p does not divide a; this special case is often called *Fermat's little theorem*.

In fact, Euler's theorem predates Lagrange's theorem by about a century. A more historically accurate proof of Euler's theorem goes like this:

Let a be an integer with $(a,n) = 1$, and consider the product $L = \prod_{1\le q<n;(q,n)=1} (aq)$. On the one hand, we have $L = a^{\Phi(n)} \prod_{1\le q<n;(q,n)=1} q$. On the other hand, for each q' with $1 \le q' < n$ such that $(q', n) = 1$, the equation $aq \equiv q' \pmod n$ has a unique solution mod n (since $(a,n) = 1$), and hence there is a unique q, with $1 \le q < n$ such that $aq \equiv q' \pmod n$.

$$L = \prod_{1\le q<n;(q,n)=1} (aq) \equiv \prod_{1\le q'<n,(q',n)=1} q' (\mathrm{mod}\ n).$$

We thus have

$$a^{\Phi(n)} \prod_{1\le q<n;(q,n)=1} q \equiv \prod_{1\le q'<n,(q',n)=1} q' (\mathrm{mod}\ n).$$

Cancelling the product from both sides (all of the factors are invertible mod n), we find $a^{\Phi(n)} \equiv 1 \pmod n$.

4. Public-Key Cryptography

4.1. **Messages as sequences of classes mod n.** Since the advent of computers, the idea of representing a message by a string of numbers is a familiar one. In practice, to do this one typically chooses a way of encoding individual characters as binary numbers of a fixed length d (usually 8 or 16 bits, i.e. binary digits.) If we then cut a message up into "blocks" of n characters and concatenate the binary representations of each character in the block, we obtain a dn-bit binary number that represents an n-character block as an integer between 0 and 2^{nd}. If we choose some modulus $m > 2^{nd}$, then we can alternatively represent a block as a class in $\mathbb{Z}/m\mathbb{Z}$. Thus we will be mainly concerned with the problem of communicating a congruence class c mod m, for some large m, between a sender A and a recipient B. The goal is to do this in such a way that any eavesdroppers on the communication can not deduce what c is (but B can).

4.2. **The RSA algorithm (Rivest–Shamir–Adleman 1978, but also Clifford Cocks 1973).** Most traditional forms of cryptography rely on a *shared secret* known to both A and B. This shared secret is effectively some invertible function f from $\mathbb{Z}/m\mathbb{Z}$ to itself. The idea is that rather than sending c to B directly, A computes $f(c)$, sends that to B, and then B computes $f^{-1}(c)$. Since eavesdroppers do not know f, they (at least in principle) can't recover c from $f(c)$.

In practice, for A and B to agree on a function f poses problems. (In particular, they have to communicate to do so, and if eavesdroppers listen to that communication they can learn f.)

The algorithm we describe today avoids this problem completely! It is what is known as a *public-key* algorithm. Instead of secrets being shared between A and B, our recipient B creates a secret *known only to B* (his private

key), and then releases additional information (the public key) to *anyone who wants to communicate with him.* For anyone to send B a message, only the public key is required, but decoding the message requires the private key.

Here is how the algorithm works. B first chooses two large prime numbers p and q (in practice, each of these is around 2^{1024} or so) and sets $m = pq$. An integer mod m thus allows you to represent 2048 bits of information, or 256 eight-bit characters. B also chooses a number e such that $(e, \Phi(m)) = 1$, and lets d be a multiplicative inverse of e mod $\Phi(m)$.

The public key, that B shares with everyone, consists of the numbers m and e.

The private part of the key, that B must keep secret, consists of the numbers p, q, and d.

To encode a message c, a sender A computes $s = c^e \pmod m$, and sends it to B.

Given an encoded message s, B decodes it by computing $s^d \pmod m$. The reason this works is that if $s = c^e$, then one has $(c^e)^d = c^{ed} = c^{1+k\Phi(m)}$, since, by construction, $ed \equiv 1 \pmod{\Phi(m)}$. Thus $s^d = c^{ed} \equiv c \pmod m$ by Euler's theorem. (Note that this works provided c is coprime to m, but it's still OK even without that, since m is squarefree - exercise using Fermat's Little Theorem plus Chinese Remainder Theorem.)

Any eavesdropper who knows c^e and wants to deduce c has to be able to compute an eth root of c mod m. *As far as we know,* this is quite difficult computationally! The best (publicly) known approaches all involve factoring m. For numbers around 2^{2048}, this is not feasible with today's computing equipment (and might well never be feasible)! On the other hand, we have no formal proof that factoring is as computationaly difficult as it seems to be. As far as I'm aware, we don't even have a formal proof that breaking RSA is as computationally difficult as factoring.

In spite of these uncertainties, our intuition and experience suggests that recovering c from c^e without knowing a factorization of m is computationally infeasible. It is this infeasibility that allows the cryptosystem to work.

4.3. **Signing with RSA.** Public-key cryptography can also be used as proof of identity. Suppose B wants to make a declaration to the world, and prove beyond all doubt that it was B who made the declaration, and not an impostor. (Perhaps this declaration is a will, or acceptance of a contract, for instance.)

B first represents the message he wants to sign as a class c mod m. To sign this class, B computes $c^d \pmod m$ using the private part of the key, and sends the world the pair (c, c^d).

Suppose A wants to verify that a pair (c, s) was a message signed by B. Then A computes $s^e \pmod m$, which requires only the public part of the key. If $s \equiv c^d \pmod m$, then $s^e = c^{de} \equiv c \bmod m$. So A just needs to check that $s^e \equiv c \pmod m$ and if so the signature is verified.

To fake a message signed by B, a forger needs to solve the problem of, given a message c, finding a signature s such that $s^e \equiv c \pmod{m}$. *This is precisely the same problem as deciphering a message sent by the algorithm above.* Thus forging signatures is just as hard as breaking the encryption!

5. A couple of other applications of Euler's theorem

We saw that if $(a, n) = 1$ then $a^{\Phi(n)} \equiv 1 \pmod{n}$. In particular, if n is prime, then $a^{n-1} \equiv 1 \pmod{n}$. Conversely, if we can find an $1 \leq a < n$ with $a^{n-1} \not\equiv 1 \pmod{n}$, then n is *not* prime. Even just taking $a = 2$ is often enough to show that n isn't prime, and using repeated squaring this can be checked quickly.

However it doesn't always work. A *Carmichael number* is a composite number such that if $(a, n) = 1$ then $a^{n-1} \equiv 1 \pmod{n}$. There are infinitely many of them (this is a hard theorem!).

Example: 561 is the smallest Carmichael number. Why does this work? $561 = 3 \cdot 11 \cdot 17$. More generally suppose that we have $n = pqr$, with p, q, r distinct primes. Then we will be OK by FLT and CRT provide that $n - 1$ is divisible by each of $p - 1$, $q - 1$, $r - 1$. But $560 = 2^4 * 5 * 7$ and $3 - 1 = 2$, $11 - 1 = 10 = 2 * 5$, $17 - 1 = 16 = 2^4$.

(Exercise: use the information below on primitive roots to show that this is the smallest Carmichael number of the form pqr.)

Another application is to showing that certain Diophantine equations have no solutions. e.g. we know that $x^2 + y^2 = a$ has no solutions if $a \equiv 3 \pmod{4}$, by looking mod 4. If we have want to do something similar mod n for some Diophantine equation, we should look for n such that $\Phi(n)$ is a small multiple of the exponents in the equation. For example, let's think about $x^5 + y^5 = a$.

We want n such that $\Phi(n)$ is divisible by 5. This suggests taking $n = 11$. Then we see that either $a^5 \equiv 0 \pmod{11}$ (if $11|a$), or $a^{10} \equiv 1 \pmod{11}$, so that $a^5 \equiv \pm 1 \pmod{11}$. So in particular $x^5 + y^5 \equiv -2, -1, 0, 1, 2 \pmod{11}$, and e.g. $x^5 + y^5 = 125$ has no solutions because $125 \equiv 4 \pmod{11}$.

6. Primitive Roots

6.1. **Basic Definitions.**

Definition 6.1. Let n be a positive integer and a an integer with $(a, n) = 1$. The *order* of a mod n is the smallest positive integer k such that $a^k \equiv 1 \pmod{n}$.

Proposition 6.2. *Let a be an integer with $(a, n) = 1$, and let k be the order of a mod n. If $a^d \equiv 1 \pmod{n}$ then k divides d.*

Proof. Write $d = qk + r$ with q, r integers and $0 \leq r < k$. Then $1 \equiv a^d = a^{qk+r} = (a^k)^q(a^r) \pmod{n}$; since $a^k \equiv 1 \pmod{n}$ it follows that $a^r \equiv 1 \bmod n$. Since $r < k$, r cannot be positive (by the definition of order), so $r = 0$ and $d = qk$. □

It now follows from Euler's theorem that for any a with $(a, n) = 1$, the order of a mod n divides $\Phi(n)$.

Definition 6.3. An integer a with $(a, n) = 1$ is a *primitive root* mod n if the order of a mod n is exactly $\Phi(n)$.

Primitive roots need not exist modulo every n. For instance, if $n = 8$, then $3, 5$, and 7 have order 2 mod 8, and 1 of course has order 1. Since $\Phi(8) = 4$, there are no primitive roots mod 8.

On the other hand, we have:

Theorem 6.4. *Let p be a prime. Then there exists a primitive root mod p.*

We will prove this shortly. In fact, this is not the strongest result possible; one has:

Theorem 6.5. *Let n be a positive integer. There exists a primitive root mod n exactly in the following cases (and no others):*

1. *$n = 1, 2$, or 4,*
2. *$n = p^r$ where p is an odd prime and $r \geq 1$, and*
3. *$n = 2p^r$ where p is an odd prime and $r \geq 1$.*

We will not prove this, although there are some special cases on the example sheets.

6.2. **Polynomials over a field.** The proof of this theorem requires some results about roots of polynomials mod p. Over the rational numbers we all know that a polynomial of degree d has at most d roots. This can fail over other rings; for instance, the polynomial $x^2 - x$ has the roots $[0]_6, [1]_6, [3]_6, [4]_6$ mod 6. The issue here is that $\mathbb{Z}/6\mathbb{Z}$ is not a field.

Lemma 6.6. *Let R be a ring, and $P(X)$ be a polynomial in X with coefficients in R. If $P(\alpha) = 0$, then there exists a polynomial $Q(X)$ with coefficients in R such that $P(X) = Q(X)(X - \alpha)$.*

Proof. We proceed by induction on the degree of P, the degree zero case being clear. Suppose the result is true for polynomials of degree $d - 1$, and let $P(X)$ have degree d. If the leading term of $P(X)$ is cX^d, let $S(X) = P(X) - cX^{d-1}(X - \alpha)$. We have $S(\alpha) = 0$, and $S(X)$ has degree $d - 1$so there exists $T(X)$ with coefficients in R such that $R(X) = T(X)(X - \alpha)$. Then $P(X) = [cX^d + T(X)](X - \alpha)$. □

Theorem 6.7. *Let F be a field, and $P(X)$ a polynomial of degree d with coefficients in F. Then $P(X)$ has at most d distinct roots in F.*

Proof. We again proceed by induction on d; the case $d = 0$ is clear. If $P(X)$ has degree d and α is a root, we can write $P(X) = (x - \alpha)Q(X)$. Now if $P(\beta) = 0$, then $(\beta - \alpha)Q(\beta) = 0$, so (since F is a field) either $\beta = \alpha$ or β is a root of $Q(X)$. By the inductive hypothesis $Q(X)$ has at most $d - 1$ roots in F, so P has at most d roots as claimed. □

As a corollary, we deduce:

Corollary 6.8. *Let p be a prime, and let d divide $p-1$. Then the polynomial $x^d - 1$ has exactly d roots mod p.*

Proof. Note that by Fermat's little theorem, $[1]_p, \ldots, [p-1]_p$ are all roots of $x^{p-1} - 1 \bmod p$. Thus $x^{p-1} - 1$ has exactly $p-1$ roots. Now fix d dividing $p-1$ and write $x^{p-1} - 1 = (x^d - 1)Q(X)$ for a polynomial $Q(X)$; $Q(X)$ has integer coefficients so we can view it as a polynomial mod p. Now $x^{p-1} - 1$ has $p-1$ roots, x^d has at most d roots, and $Q(X)$ has at most $p - 1 - d$ roots. We must thus have equality; i.e. $x^d - 1$ has d roots mod p. □

Another way of stating the corollary is to say that for any d dividing $p-1$, there are exactly d elements of $(\mathbb{Z}/p\mathbb{Z})^\times$ whose order divides d.

6.3. **Existence of primitive roots mod p.** We are now ready to prove the existence of primitive roots mod p. We first recall from the first example sheet that we have:

$$\sum_{d|n;d>0} \Phi(d) = n.$$

Theorem 6.9. *Let p be a prime. Then for any d dividing $p-1$, there are exactly $\Phi(d)$ elements of order d in $(\mathbb{Z}/p\mathbb{Z})^\times$. In particular there are $\Phi(p-1)$ primitive roots mod p.*

Proof. We prove this by strong induction on d; the case $d = 1$ is clear. Fix d. The inductive hypothesis tells us that for any d' dividing d and strictly less than d there are $\Phi(d')$ elements of exact order d'. On the other hand there are a total of d elements of $(\mathbb{Z}/p\mathbb{Z})^\times$ of order dividing d. Thus the number of elements of $(\mathbb{Z}/p\mathbb{Z})^\times$ of order exactly d is:

$$d - \sum_{d'|d, 0<d'<d} \Phi(d')$$

and this is precisely $\Phi(d)$. □

6.4. **Discrete logarithms.** If g is a primitive root mod n then, by definition, the order of g is $\Phi(n)$. It follows that the powers $g, g^2, \ldots g^{\Phi(n)}$ are all distinct mod n (since if $g^i \equiv g^j \bmod n$ for $0 < i < j < n$ we would have $g^{j-1} \equiv 1 \pmod n$.) Since there are only $\Phi(n)$ classes in $(\mathbb{Z}/n\mathbb{Z})^\times$, this says that if g is a primitive root then every class in $(\mathbb{Z}/n\mathbb{Z})^\times$ is a power of g. In other words, the group $(\mathbb{Z}/n\mathbb{Z})^\times$ is a cyclic group of order $\Phi(n)$, generated by g. For this reason primitive roots are often called *generators*.

Another way to say this is that when a primitive root g exists mod n, the map: $\mathbb{Z}/\Phi(n)\mathbb{Z} \to (\mathbb{Z}/n\mathbb{Z})^\times$ taking $[a]_{\Phi(n)}$ to $[g^a]_n$ is an isomorphism, from the additive group of $\mathbb{Z}/\Phi(n)\mathbb{Z}$ to $(\mathbb{Z}/n\mathbb{Z})^\times$. It thus has an inverse, which we call the *discrete logarithm to the base g*.

Explicitly, if g is a primitive root mod n, then the discrete logarithm to the base g (denoted $\log_g$) is defined by $\log_g a = [k]_{\Phi(n)}$, where k is an integer such that $g^k \equiv a \pmod n$.

One use of the discrete logarithm is to solve exponential equations mod n, if n admits a primitive root g. For instance, by applying $\log_g$ to both sides of the equation $x^d \equiv a \pmod{n}$, we obtain the linear congruence equation $d \log_g x \equiv \log_g a \pmod{\Phi(n)}$ and we can solve those with techniques explained earlier.

It is expect that the discrete logarithm is hard to compute (much in the way that it is expected to be hard to crack RSA, but we don't know for sure). Here's a practical application of this: safely storing passwords. Let p be a big prime, big enough so that all passwords can be thought of as residues mod $(p-1)$, and fix a primitive root g mod p. Then if someone sets their password to be a, you can compute and store g^a. If they want to login with b, you compute g^b, and check if it equals what you stored.

Even if someone has access to what you've stored, and to g, they still can't recover the password a without solving the discrete logarithm problem. (Of course, nor can you, so it's not so good if you require people to be able to be reminded of their passwords!)

7. Quadratic Reciprocity

7.1. **Quadratic Residues.**

Definition 7.1. Let p be an odd prime and a an integer not divisible by p. We say that a is a *quadratic residue mod p* if there exists an integer d with $d^2 \equiv a \pmod{p}$. If no such d exists, we say that a is a *quadratic non-residue mod p.*

Note that, by convention, integers a divisible by p are neither quadratic residues nor quadratic non-residues mod p.

Lemma 7.2. *If $p > 2$ then there are exactly $(p-1)/2$ quadratic residues mod p, and $(p-1)/2$ quadratic non-residues.*

Proof. The quadratic residues are the image of the homomorphism $(\mathbb{Z}/p\mathbb{Z})^\times \to (\mathbb{Z}/p\mathbb{Z})^\times$, $x \mapsto x^2$. This has kernel $x = \pm 1$, so the image has order $(p-1)/2$. In a more down to earth language, we can consider the squares $1^2, 2^2, \dots, (p-1)^2$; we have $x^2 \equiv y^2 \pmod{p}$ if and only if $x \equiv \pm y \pmod{p}$, and we never have $x \equiv x \pmod{p}$. □

Note that of course if $p = 2$ then 1 is a quadratic residue.

Proposition 7.3. *Let a, b be integers with $(a,p) = (b,p) = 1$. Then:*

- *If a and b are quadratic residues mod p, then so is ab.*
- *If a is a quadratic residue mod p and b is a quadratic non-residue mod p, then ab is a quadratic non-residue mod p.*
- *If a and b are quadratic non-residues mod p, then ab is a quadratic residue mod p.*

Proof. The shortest proof is to observe that if $H \leq (\mathbb{Z}/p\mathbb{Z})^\times$ is the subgroup of quadratic residues (it's a subgroup as it's the image of a homomorphism), then $G/H \equiv \mathbb{Z}/2\mathbb{Z}$ (the only possible group of order 2).

More down to earth: for the first claim, if $a \equiv d^2 \pmod p$ and $b \equiv c^2 \pmod p$, then $ab \equiv (cd)^2 \pmod p$. For the second, if $a \equiv d^2 \pmod p$ and $ab \equiv c^2 \pmod p$, then we have $[b]_p = [a]_p^{-1}[ab]_p = ([d]_p^{-1}[c]_p)^2$. For the third, let a be a quadratic non-residue. Then the set of values $[a]_p[x]_p$, $1 \leq x \leq p-1$, is a complete set of residues for $(\mathbb{Z}/p\mathbb{Z})^\times$, and we have seen that if x is a quadratic residue, then this gives a quadratic non-residue. But there are exactly $(p-1)/2$ quadratic residues, and exactly $(p-1)/2$ quadratic residues, so we're done by counting.

Alternative proof of the third part: choose a primitive root g mod p and write $a \equiv g^r \pmod p$, $b \equiv g^s \pmod p$. Then r and s are odd since a and b are non-residues. But then $ab \equiv g^{r+s} \pmod p$ and $r+s$ is even. □

Definition 7.4. The Legendre symbol $\left(\frac{a}{p}\right)$, for p prime and a an integer, is defined by:

- $\left(\frac{a}{p}\right) = 1$ if a is a quadratic residue mod p,
- $\left(\frac{a}{p}\right) = -1$ if a is a quadratic non-residue mod p,
- $\left(\frac{a}{p}\right) = 0$ if $p \mid a$.

The proposition above then amounts to saying that the map: $(\mathbb{Z}/p\mathbb{Z})^\times \to \pm 1$ defined by $[a]_p \mapsto \left(\frac{a}{p}\right)$ is a group homomorphism.

In fact, the existence of primitive roots gives us an easy description of this map:

Theorem 7.5 (Euler's Criterion)**.** *Let p be an odd prime, and a an integer not divisible by p. Then*

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

Proof. If p divides a this is clear. Otherwise, note that $(a^{\frac{p-1}{2}})^2 \equiv 1 \pmod p$ by Fermat's little theorem. Thus (since the polynomial $x^2 - 1$ has only two roots mod p) we have $a^{\frac{p-1}{2}} \equiv \pm 1 \pmod p$. If a is a quadratic residue mod p then $a \equiv d^2 \pmod p$ for some d, and then $a^{\frac{p-1}{2}} \equiv d^{p-1} \equiv 1 \pmod p$. But we saw above that $x^{(p-1)/2} - 1$ has exactly (or even just at most) $(p-1)/2$ roots mod p, so we're done.

Alternatively, if a is a quadratic non-residue, then $a \equiv g^r \pmod p$ for some odd integer r and primitive root g mod p. Then $a^{\frac{p-1}{2}} \equiv g^{\frac{r(p-1)}{2}}$. Since r is odd, $\frac{r(p-1)}{2}$ is not divisible by $p-1$, so (because g has order $p-1$) $g^{\frac{r(p-1)}{2}}$ is not congruent to 1 mod p. It must therefore be -1 mod p. □

7.2. **Computing Legendre symbols.** Euler's criterion lets us determine, for fixed p, which a are quadratic residues mod p. What if we fix a, and ask for which odd primes p is a a quadratic residue?

When $a = -1$, Euler's criterion gives an easy answer: $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$, so -1 is a quadratic residue mod p if, and only if $\frac{p-1}{2}$ is even. In other words:

Proposition 7.6. *The integer -1 is a quadratic residue mod p if p is 1 mod 4, and a quadratic non-residue if p is 3 mod 4.*

For example, if $p = 5$ then it is a square, and if $p = 7$ it isn't, and in each case we can check directly.

When $a = 2$, the situation is more difficult, but still amenable to a direct approach:

Proposition 7.7 (A special case of Gauss' Lemma)**.** *:*

- $\left(\frac{2}{p}\right) = 1$ *if* $p \equiv 1, 7$ *(mod 8)*
- $\left(\frac{2}{p}\right) = -1$ *if* $p \equiv 3, 5$ *(mod 8).*

Proof. Let $q = \frac{p-1}{2}$, and set $Q = 2(4)(6)(8)\dots(p-1) = 2^q(q!)$. Reduce all the factors in the product defining Q mod p so that they lie between $-q$ and q (i.e., subtract p from every factor greater than q.) Let Q' be the resulting product $2(4)(8)\dots(-3)(-1)$. We have $Q' \equiv Q \pmod{p}$. On the other hand, the factors in the product defining Q' are the even integers from 1 to q and the negatives of the odd integers from 1 to q. Thus $Q' = (-1)^r q!$, where r is the number of odd integers between 1 and q. We thus have $2^q(q!) \equiv (-1)^r(q!) \pmod{p}$, so $2^q \equiv (-1)^r \pmod{p}$. The result follows by noting that r is even precisely when p is 1 or 7 mod 8, and invoking Euler's criterion.

More precisely, if $p = 8k+1$ then we are counting odd integers in $[1, 4k]$, so $r = 2k$; if $p = 8k+3$ then we are counting odd integers in $[1, 4k+1]$, so $r = 2k+1$; if $p = 8k+5$ then we are counting odd integers in $[1, 4k+2]$, so $r = 2k+1$; and $p = 8k+7$ then we are counting odd integers in $[1, 4k+3]$, so $r = 2k+2$. □

Since we have $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$, to answer this question in full generality it suffices to answer it for $a = -1$, for $a = 2$, and for a an odd prime. In the latter case we have:

Theorem 7.8 (Law of Quadratic Reciprocity)**.** *Let p and q be odd primes. Then:*

- $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$ *if either p or q is 1 mod 4;*
- $\left(\frac{p}{q}\right) = -\left(\frac{q}{p}\right)$ *if both p and q are 3 mod 4.*

One can rephrase this a bit more tersely as the equivalent statement:

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

Note that this implies that for each odd prime q, the question of whether q is a quadratic residue mod p has an answer in terms of congruence conditions mod q and mod 4. From this and the Chinese remainder theorem, we can deduce that the question: "for which primes p is a a quadratic residue mod p?" has an answer in terms of congruence conditions on p.

For example, if $p \neq 5$ is an odd prime, then we see that 5 is a quadratic residue modulo p if and only if p is a quadratic residue mod 5, i.e. if and only if $p \equiv \pm 1 \pmod 5$.

Slightly more complicated example, let $p \neq 7$ be an odd prime. When is 7 a quadratic residue modulo p? Well, if $p \equiv 1 \pmod 4$ then this is if and only if p is a quadratic residue mod 7, so if and only if $p \equiv 1, 2, 4 \pmod 7$. If $p \equiv -1 \pmod 4$ then if and only if $p \equiv 3, 5, 6 \pmod 7$. Putting this together with the Chinese Remainder Theorem, we see that 7 is a quadratic residue modulo p if and only if $p \equiv 1, 3, 9, 19, 25, 27 \pmod{28}$

More generally, one can ask, given a monic polynomial f with integer coefficients, for which primes p does f have a root? (The above case is the case of the polynomial $X^2 - a$.) This is a very deep question in number theory! Indeed, we are still extremely far from having a complete answer. One question it is natural to ask is: for which f does the above question have an answer given in terms of congruence conditions on p. A deep branch of algebraic number theory called *class field theory* tells us that this will happen precisely when the field extension determined by f has abelian Galois group. Beyond this we know very little, but there are connections to the theory of modular forms.

7.3. **Proof of Quadratic Reciprocity.** Quadratic Reciprocity was one of the deepest results of the 18th century, and there are many approaches to proving it, none of which are particularly simple. The more motivated ones require algebraic number theory, and even then the motivation is really coming from class field theory, which is a long way beyond the boundaries of this course.

The proof that we give is due to Rousseau, from 1991 (!). It has the merits of being elementary and relatively easy to remember, and of resembling the proof of Gauss' Lemma that we gave above.

Let p, q be distinct odd primes, and consider the group

$$(\mathbb{Z}/pq\mathbb{Z})^\times \cong (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times.$$

What we are going to do is to compare the products of different sets of coset representatives for the subgroup $\{\pm 1\}$; that is, we will look at different ways of choosing exactly one element of each pair $\{x, -x\}$. We'll always write everything as a pair (x, y) where $x \in (\mathbb{Z}/p\mathbb{Z})^\times$, $y \in (\mathbb{Z}/q\mathbb{Z})^\times$.

Firstly we recall:

Theorem 7.9 (Wilson's Theorem)**.** *If p is prime then* $(p-1)! \equiv -1 \pmod p$.

Proof. Equivalently, show that $(p-2)! \equiv 1 \pmod{p}$, and that follows as it's a product of elements x, x^{-1} with $x^{-1} \neq x$. □

Write $P = (p-1)/2$, $Q = (q-1)/2$. As our first set of coset representatives, consider the product of all pairs $(x, y) \in (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times$ with $1 \leq x \leq (p-1)/2$, $1 \leq y \leq q-1$. Then the product of the y-coordinates is $(q-1)!^P \cong (-1)^P \pmod{q}$. The product of the x-coordinates is $P!^{q-1}$ in the same way, so

$$A = (P!^{q-1}, (-1)^P).$$

Similarly we let B be the product of the pairs (x, y) with $1 \leq x \leq p-1$, $1 \leq y \leq (q-1)/2$. In the same way, we get

$$B = ((-1)^Q, Q!^{p-1}).$$

Finally, let C be the product of the representatives in $\mathbb{Z}/pq\mathbb{Z}$ given by $1 \leq i \leq (pq-1)/2$, $(i, pq) = 1$. Let's figure out the product of the x-coordinates; it's

$$\begin{aligned}
\prod_{1 \leq i \leq (pq-1)/2, (i,pq)=1} i &= \Big(\prod_{1 \leq i \leq (pq-1)/2, (i,p)=1} i\Big)\Big(\prod_{1 \leq i \leq (pq-1)/2, (i,p)=1, q|i} i\Big)^{-1} \\
&= \Big(\prod_{1 \leq i \leq (pq-1)/2, (i,p)=1} i\Big)\Big(\prod_{1 \leq i \leq (p-1)/2,} (qi)\Big)^{-1} \\
&= \Big(\prod_{1 \leq i \leq p(q-1)/2+(p-1)/2, (i,p)=1} i\Big)/P!q^P \\
&= (p-1)!^Q P!/P!q^P \\
&= (-1)^Q/q^P.
\end{aligned}$$

So by symmetry we have

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

by Euler's criterion. Now, we can compare A, B, C. We know that they agree up to possibly multiplying by the pair $(-1, -1) \in (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times$. Comparing the second coordinates, we have $C = \left(\frac{p}{q}\right) A$ and comparing first coordinates, $C = \left(\frac{q}{p}\right) B$. So

$$B = \left(\frac{p}{q}\right)\left(\frac{q}{p}\right) A.$$

But we can compare A, B more directly: to go between A, B we need to swap the signs of the PQ elements of the form (x, y) with $1 \leq x \leq P$ and $Q < y \leq q-1$. So we have

$$B = (-1)^{PQ} A,$$

and comparing these two expressions for B/A, we have proved quadratic reciprocity.

8. Sums of Two Squares

8.1. **Sums of two squares.** Here we begin to answer the question: which integers are representable as a sum of two squares?

Definition 8.1. We will say that a positive integer n is a *sum of two squares* if there exist integers x and y such that $n = x^2 + y^2$.

Of course there is an algorithm to work this out, which is just to use brute force, as $|x|$ and $|y|$ are both at most $\sqrt{n}$.

In a negative direction, note that any square is congruent to zero or one mod 4, so if n is a sum of two squares, then n cannot be congruent to 3 mod 4.

On the other hand, we have:

Lemma 8.2. *If m and n are sums of two squares, then so is mn.*

Proof. If $m = x^2 + y^2$, and $n = u^2 + v^2$, then $mn = (x^2 + y^2)(u^2 + v^2) = (xu - yv)^2 + (xv + yu)^2$. □

We will later see a more conceptual interpretation of this argument.

In light of this result it makes sense to focus first on which primes are the sum of two squares. Indeed, we have:

Theorem 8.3 ((Fermat's Two-Square Theorem))**.** *Every prime congruent to 1 mod 4 is the sum of two squares.*

We will prove this in the next section. For now, we note a consequence. Let $\mathrm{ord}_p(n)$ be the largest a such that p^a divides n.

Corollary 8.4. *Let n be a positive integer, and suppose that for each prime p congruent to 3 (mod 4), we have that $\mathrm{ord}_p(n)$ is even. Then n is the sum of two squares.*

Proof. Such an n can be written as the product of a power of 2, powers of primes p congruent to 1 mod 4, and powers of q^2 for q congruent to 3 mod 4. But $2 = 1^2 + 1^2$, every p congruent to 1 mod 4 is a sum of two squares by the theorem, and $q^2 = q^2 + 0^2$ is a sum of two squares. Thus n is a product of sums of two squares and is therefore itself a sum of two squares. □

In fact it will turn out that these are *precisely* the n which can be expressed as the sum of two squares, but we will only prove this later, when we have more tools.

8.2. **Proof of the two square theorem.** We first note that if p is a prime congruent to one mod 4, then Euler's criterion tells us that -1 is a quadratic residue mod p. We thus have an n, which we can take between 0 and $p-1$, such that $n^2 \equiv -1 \pmod{p}$. In particular p divides n^2+1, so we can write $n^2+1 = pr$, with $1 \leq r < p$. Note that if $r = 1$ then we are done, as then $p = n^2+1$.

The strategy of the proof will be to inductively reduce the r appearing above. We will use the following proposition to do so:

Proposition 8.5 ((Fermat Descent))**.** *Suppose we have $a^2+b^2 = pr$, with a, b integers and $1 < r < p$. Then there exists $1 \leq r' < r$ and integers x, y such that $x^2+y^2 = pr'$.*

Proof. Choose u, v such that $u \equiv a \pmod{r}$, $v \equiv b \pmod{r}$, and $-\frac{r}{2} \leq u, v \leq \frac{r}{2}$. Then $u^2+v^2 \equiv a^2+b^2 \equiv 0 \pmod{r}$, so we can write $u^2+v^2 = rr'$. Since u^2 and v^2 are at most $\frac{r^2}{4}$, we have that $r' \leq \frac{r}{2}$. On the other hand, we cannot have $r' = 0$, as then $u = v = 0$, so r divides both a and b and then r^2 divides $a^2+b^2 = pr$. It would then follow that r divides p, which is impossible. Thus $1 \leq r' < r$.

Now $(a^2+b^2)(u^2+v^2) = r'r^2p$. Rewriting this as a sum of two squares, we have: $(au+bv)^2 + (av-bu)^2 = r'r^2p$. Set $x = \frac{au+bv}{r}$, $y = \frac{av-bu}{r}$; then $x^2+y^2 = r'p$. We must check that x and y are integers, however.

But $au+bv \equiv a^2+b^2 \equiv 0 \pmod{r}$, and $av-bu \equiv ab-ba \equiv 0 \pmod{r}$, so this is clear. □

Now to express p as a sum of two squares, set $a_0 = n, b_0 = 1, r_0 = r$, and repeatedly apply the proposition to get a_i, b_i, r_i with $a_i^2+b_i^2 = r_ip$ and the r_i decreasing, but always at least one. Eventually one has $r_m = 1$ for some m and the claim follows.

9. Quadratic Rings and Euclidean Domains

9.1. The Gaussian Integers.

Definition 9.1. The ring of *Gaussian integers*, denoted $\mathbb{Z}[i]$, is the subring of $\mathbb{C}$ consisting of all complex numbers of the form $a+bi$, where a and b are integers.

To see that $\mathbb{Z}[i]$ is a subring one must of course check that it is closed under addition and multiplication; this is easy.

There is a natural map $N : \mathbb{Z}[i] \to \mathbb{Z}_{\geq 0}$ defined by $N(z) = z\overline{z}$. We have $N(a+bi) = a^2+b^2$. On the other hand, since $\overline{zw} = \overline{z}\,\overline{w}$, we have $N(zw) = N(z)N(w)$.

Let $z = a+bi$, $w = c+di$. We then have $zw = (ac-bd)+(ad+bc)i$. Applying the formula $N(zw) = N(z)N(w)$ we find that $(ac-bd)^2+(ad+bc)^2 = (a^2+b^2)(c^2+d^2)$. This gives a conceptual reason for our earlier observation that the product of sums of two squares is a sum of two squares.

9.2. **The ring $\mathbb{Z}[\alpha]$.** This suggests that we can deduce interesting results about $\mathbb{Z}$ by considering larger subrings of $\mathbb{C}$. One way to obtain such subrings is to start from $\mathbb{Z}$ and add an extra element α in $\mathbb{C}$. This is slightly more subtle for a general α than it is for $\alpha = i$.

Definition 9.2. Let $\alpha \in \mathbb{C}$. Then the ring $\mathbb{Z}[\alpha]$ is the smallest subring of $\mathbb{C}$ containing α. [n.b. for this course, a ring always contains 1.]

As usual one has to check that this makes sense. An alternative definition is to take $\mathbb{Z}[\alpha]$ to be the intersection of all subrings of $\mathbb{C}$ containing α. This intersection clearly contains 0, 1 and α, so it is nonempty. It's also closed under addition and multiplication, since it's an intersection of sets that are closed under these operations. Therefore it's a subring of $\mathbb{C}$ that, by construction is contained in any subring of $\mathbb{C}$ that contains α.

Let's show that for $\alpha = i$, this agrees with our earlier definition of $\mathbb{Z}[i]$. We've already shown that the set of all integers of the form $a+bi$ is a subring of $\mathbb{C}$. On the other hand, any subring of $\mathbb{C}$ containing i is closed under addition and multiplication, and contains 1, so it contains every complex number of the form $a + bi$. Thus our new definition is consistent with our old one.

Note that for arbitrary α, it is *not necessarily true* that $\mathbb{Z}[\alpha]$ consists of all complex numbers of the form $a + b\alpha$ for a and b integers (although it always contains all complex numbers of that form). To see this, consider examples like $\mathbb{Z}[\pi]$, $\mathbb{Z}[\frac{1}{p}]$ for p some prime, or $\mathbb{Z}[\beta]$ where β is a cube root of 2. (For example, the last ring contains β^2, which is not of the form $a + b\beta$ for any integers a and b.)

9.3. **Quadratic subrings of $\mathbb{C}$.**

Definition 9.3. An element α of $\mathbb{C}$ is an *algebraic integer of degree two* (alternatively, a *quadratic algebraic integer*) if there exists a polynomial of the form $P(X) = X^2 + aX + b$ with a, b integers such that $P(X)$ has no rational (equivalently, integer) roots and $P(\alpha) = 0$.

For instance, i is an algebraic integer of degree two since $(i)^2 + 1 = 0$.

Suppose that α is an algebraic integer of degree two, and let a, b be integers such that $\alpha^2 + a\alpha + b = 0$. Then, for x, y, z, w integers, we have

$$(x+y\alpha)(z+w\alpha) = xz+(xw+yz)\alpha+yw\alpha^2 = (xz-byw)+(xw+yz-ayw)\alpha.$$

In particular the set of complex number of the form $x + y\alpha$ (x,y integers) is closed under addition and multiplication and is therefore a subring of $\mathbb{C}$. Since this subring contains α, it contains $\mathbb{Z}[\alpha]$. On the other hand it is clear that this subring is contained in $\mathbb{Z}[\alpha]$, so the two must be equal. We thus have:

Proposition 9.4. *If α is an algebraic integer of degree two, then $\mathbb{Z}[\alpha]$ is equal to the set of complex numbers of the form $x + y\alpha$, where x and y are integers.*

For α an algebraic integer of degree two, we will say that $\mathbb{Z}[\alpha]$ is a *real quadratic subring of* $\mathbb{C}$ if α is a real number, and an *imaginary quadratic subring of* $\mathbb{C}$ if α is not real.

If $\mathbb{Z}[\alpha]$ is imaginary quadratic, then, as with $\mathbb{Z}[i]$, we can define a *norm* $N : \mathbb{Z}[\alpha] \to \mathbb{Z}_{\geq 0}$ by setting $N(z) = z\overline{z}$. Note that $\overline{\alpha}$ is also a root of $\alpha^2 + a\alpha + b$ in this case, so we have $\alpha + \overline{\alpha} = -a$, $\alpha\overline{\alpha} = b$. Thus we have $N(x + y\alpha) = (x + y\alpha)(x + y\overline{\alpha}) = x^2 + xy(\alpha + \overline{\alpha}) + y^2\alpha\overline{\alpha} = x^2 - axy + by^2$. Note that this is multiplicative: $N(zw) = N(z)N(w)$.

If $\mathbb{Z}[\alpha]$ is real quadratic, we let α^* denote the root of $X^2 + aX + b$ that is not equal to α (note that this is no longer equal to $\overline{\alpha}$.) In this case we define $N(x + y\alpha) = (x + y\alpha)(x + y\alpha^*) = x^2 - axy + by^2$. We thus get a map $N : \mathbb{Z}[\alpha] \to \mathbb{Z}$. This is again multiplicative, but no longer nonnegative. For instance, in $\mathbb{Z}[\sqrt{2}]$, $N(1 + \sqrt{2}) = (1 + \sqrt{2})(1 - \sqrt{2}) = -1$.

If $z = a + b\alpha \in \mathbb{Z}[\alpha]$ we write $z^* = a + b\alpha^*$, so that $N(z) = zz^*$.

9.4. **Factorization in quadratic rings.** We can study factorization in quadratic rings in a way analogous to the situation over $\mathbb{Z}$. First, some definitions:

Definition 9.5. An element of $\mathbb{Z}[\alpha]$ is a *unit* if it has a multiplicative inverse; that is, if it lies in $\mathbb{Z}[\alpha]^\times$. Two elements z, w of $\mathbb{Z}[\alpha]$ are *associates* if there exists a unit $u \in \mathbb{Z}[\alpha]^\times$ such that $z = uw$.

Note that if $u \in \mathbb{Z}[\alpha]$ is a unit, there exists v such that $uv = 1$. Taking norms we find that $N(u) = \pm 1$ (if $\mathbb{Z}[\alpha]$ is imaginary quadratic, then this means $N(u) = 1$.) Conversely, if $N(u) = \pm 1$, then $uu^* = \pm 1$, so either u^* or $-u^*$ is a multiplicative inverse of u.

We thus see, for instance, that the units in $\mathbb{Z}[i]$ are $\pm 1, \pm i$.

Definition 9.6. If z, u are elements of $\mathbb{Z}[\alpha]$, we say $z \mid u$ if there exists $r \in \mathbb{Z}[\alpha]$ such that $u = zr$.

Note that every element of $\mathbb{Z}[\alpha]$ is divisible by units and its associates.

Definition 9.7. A nonzero element z of $\mathbb{Z}[\alpha]$ is *irreducible* if its only divisors are units and its associates.

Note that for the ring $\mathbb{Z}$, we called such elements prime. Here we will use the word irreducible instead, and save the word prime for a stronger condition (that is equivalent to irreducibility in the ring $\mathbb{Z}$).

e.g. in $\mathbb{Z}[i]$, using the norm we can check that 3 is irreducible, but 5 isn't.

It is also possible to definite greatest common divisors over $\mathbb{Z}[\alpha]$, although there is no longer any ordering on $\mathbb{Z}[\alpha]$. We thus need to take a different approach to defining greatest common divisors:

Definition 9.8. Let z and w be elements of $\mathbb{Z}[\alpha]$, not both zero. An element r of $\mathbb{Z}[\alpha]$ is a *greatest common divisor* of z and w if:

- r divides both z and w, and
- if $s \in \mathbb{Z}[\alpha]$ divides both z and w, then s divides r.

You showed in the exercises that over $\mathbb{Z}$, the greatest common divisor had this property, and that this property characterized the greatest common divisor up to sign. It's thus sensible to take this a definition over $\mathbb{Z}[\alpha]$. Note, however, that it is not clear from this definition that greatest common divisors always exist. They are also not unique: if a greatest common divisor does exist, then any associate is a greatest common divisor as well. On the other hand any two greatest common divisors are associates.

A natural question to ask is whether factorizations into irreducibles are unique in $\mathbb{Z}[\alpha]$. In fact, it is not hard to show that factorizations into irreducibles exist:

Proposition 9.9. *Let z be an element of $\mathbb{Z}[\alpha]$. Then z factors into irreducibles.*

Proof. Suppose there is an element of $\mathbb{Z}[\alpha]$ that does not admit a factorization into irreducibles. Then we can find such an element z such that $|N(z)|$ is minimal. Note that z itself cannot be irreducible, so we have $z = uv$ for elements u, v of $\mathbb{Z}[\alpha]$, neither of which is a unit. We have $|N(z)| = |N(u)||N(v)|$ with neither $|N(u)|$ or $|N(v)|$ equal to 1. Thus $|N(u)|, |N(v)| < |N(z)|$, so (by minimality) u and v admit factorizations into irreducibles. But then z does as well. □

On the other hand, it is *not* true, for an arbitrary $\mathbb{Z}[\alpha]$, that such factorizations are unique. For instance, in $\mathbb{Z}[\sqrt{-5}]$, one has $6 = (1+\sqrt{-5})(1-\sqrt{-5}) = 2 \cdot 3$, and it is not hard to see (for instance, by considering the norms of possible divisors) that 2, 3, $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$ are all irreducible, and none are associates of each other.

9.5. **Euclidean domains.** In certain cases, however, unique factorization *does* hold. One way to prove this, is, when possible, to mimic the proof of unique factorization for $\mathbb{Z}$. To do this one needs a version of Euclid's algorithm.

Definition 9.10. A ring R is a *Euclidean domain* if it is an integral domain, and there exists a function $N : R \to \mathbb{Z}_{\geq 0}$ such that for all $a, b \in R$, with $b \neq 0$, there exist $q, r \in R$ such that

- $a = qb + r$, and
- either $r = 0$, or $N(r) < N(b)$.

If R is a Euclidean domain, then given a pair $a = a_0, b = b_0$, with b_0 nonzero, one can form a sequence of pairs a_i, b_i by setting $a_{i-1} = q_{i-1}b_{i-1} + r_{i-1}$, with $N(r_{i-1}) < N(b_{i-1})$, and taking $a_i = b_{i-1}$, $b_i = r_{i-1}$. Just as over the integers, this terminates with $b_r = 0$ for some r. Then one shows that a_r divides a_i and b_i for all i, and that a_r is a linear combination of a_i and b_i for all i, exactly as for $\mathbb{Z}$.

One thus has:

Theorem 9.11. *Let R be a Euclidean domain, and let a and b be elements of R, with b nonzero. Then there exist elements x and y of r such that $xa + yb$ is a greatest common divisor of a and b.*

Proof. The Euclidean algorithm gives an a_r, x, y in R such that $a_r = xa+yb$ and a_r divides both a and b. It suffices to show that a_r is a greatest common divisor of a and b. To see this, suppose that s divides a and b. Then s divides $xa + yb$, so s divides a_r. Thus a_r has both properties required of a greatest common divisor. □

As one might expect from the argument over $\mathbb{Z}$, this allows us to show:

Theorem 9.12. *Let r, a, b in R, with R a Euclidean domain and r irreducible. Then if r divides ab, either r divides a or r divides b.*

Proof. Suppose r does not divide a. Then the only common divisors of r and a are units, so we can write $1 = ax + ry$. Then $b = abx + rby$, and both abx and rby are divisible by r, so r divides b. □

Definition 9.13. An element r of a ring R is *prime* if for any a, b in R such that r divides ab, either r divides a or r divides b.

The above theorem amounts to the statement that irreducible elements of a Euclidean domain are prime. From this one can conclude, as one does over $\mathbb{Z}$, that any two factorizations of an element into irreducibles coincide up to permutation and replacing by associates:

Theorem 9.14. *Let R be a Euclidean domain and let $p_1 \dots p_r = q_1 \dots q_s = x$ be two factorizations of $x \in R$ into irreducibles. Then $s = r$ and one may rearrange the q's so that q_i is an associate of p_i for all i.*

The proof is identical to the case $R = \mathbb{Z}$.

This is all well and good, but it would be better if we had examples of Euclidean domains. Obviously $\mathbb{Z}$ is a Euclidean domain, but we knew all this about $\mathbb{Z}$ already. On the other hand, we have:

Proposition 9.15. *The ring $\mathbb{Z}[i]$ is a Euclidean domain.*

Proof. We must show there is a function N on $\mathbb{Z}[i]$ with the appropriate properties; as our choice of notation suggests, we will take this to be the usual norm $N(z) = z\overline{z}$. Let a and b be elements of $\mathbb{Z}[i]$ with b nonzero. Let $q' = \frac{a}{b}$; this is an element of $\mathbb{C}$, and we can write $q' = x + yi$ with x, y real. Set $q = c+di$, where c and d are integers such that $|x-c| \leq \frac{1}{2}$ and $|y-d| \leq \frac{1}{2}$. Then $N(q - q') \leq \frac{1}{2}$. Set $r = a - bq$. Then we have $N(\frac{a}{b} - q)N(b) \leq \frac{N(b)}{2}$, so $N(r) = N(a - bq) \leq \frac{N(b)}{2}$. Thus N is a Euclidean norm and the claim follows. □

As a consequence we deduce that every nonzero element of $\mathbb{Z}[i]$ admits a unique factorization into irreducibles.

On the other hand, $\mathbb{Z}[i]$ is not the only quadratic ring in which unique factorization holds. They seem to be relatively rare in general, however. For instance, there are only nine imaginary quadratic rings in which unique factorization holds: the rings $\mathbb{Z}[i]$, $\mathbb{Z}[\sqrt{-2}]$, $\mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$, $\mathbb{Z}[\frac{1+\sqrt{-7}}{2}]$, $\mathbb{Z}[\frac{1+\sqrt{-11}}{2}]$, $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$, $\mathbb{Z}[\frac{1+\sqrt{-43}}{2}]$, $\mathbb{Z}[\frac{1+\sqrt{-67}}{2}]$, and $\mathbb{Z}[\frac{1+\sqrt{-163}}{2}]$. It's easy to show this for the first five of these (they are Euclidean) and somewhat more difficult to show for the last four. The fact that these are the only imaginary quadratic rings where one has unique factorization is due to Heegner, and wasn't established till the 1950's.

By contrast, it is not known whether there are infinitely many real quadratic rings with unique factorization, although it is expected that there are, and also believed that they are all Euclidean. However, they can be Euclidean without being norm-Euclidean, and indeed there are known to be only finitely many norm-Euclidean examples. A specific example is that $Z[(1+\sqrt{69})/2]$ is Euclidean but not norm-Euclidean.

10. Sums of Two squares revisited

10.1. **Sums of two squares.** Note that the norm in $\mathbb{Z}[i]$ is given by $N(a+bi) = a^2+b^2$. We can thus rephrase the question of which integers are representable as a sum of two squares by asking "which integers are norms in $\mathbb{Z}[i]$?". Since the norm is multiplicative, and every element of $\mathbb{Z}[i]$ is a product of primes, we can answer this question by asking "what are the primes in $\mathbb{Z}[i]$, and what are their norms?

Let's try to figure out the primes of $\mathbb{Z}[i]$.

Lemma 10.1. *Let p be a prime in $\mathbb{Z}[i]$. Then there exists an integer prime q such that either $N(p) = q$ or $N(p) = q^2$. In the latter case p is an associate of q. Moreover, $N(p) = q$ if, and only if, q is the sum of two squares.*

Proof. Let $n = N(p)$, and factor $n = q_1 q_2 \dots q_r$ as a product of integer primes. Since $n = p\overline{p}$, we have that p divides $q_1 \dots q_r$, so (since p is prime), p divides q_i for some i. Let $q = q_i$. We have $q = px$. Then $q^2 = N(q) = N(p)N(x)$, so $N(p)$ divides q^2. Since $N(p)$ is not one (if it were p would be a unit), we then have either $N(p) = q$ or $N(p) = q^2$. First suppose that $N(p) = q^2$. Since p divides q we have $pu = q$; since $N(p) = N(q) = q^2$ we must have $N(u) = q$, so u is a unit and p is an associate of q.

Suppose $N(p) = q$. Writing $p = a+bi$ we see that $q = a^2+b^2$. Conversely, if $q = a^2+b^2$, then $q = (a+bi)(a-bi)$. Since p divides q we have either p divides $a+bi$ or p divides $a-bi$; in either case $N(p)$ divides q and must thus equal q. □

Corollary 10.2. *The primes in $\mathbb{Z}[i]$ are either of the form $a+bi$, where a^2+b^2 is an integer prime, or q, where q is an integer prime that is not the sum of two squares.*

We have thus reduced the whole question to the question of which primes are the sum of two squares, which we have already answered. However,

our new perspective also gives us an alternative approach to this question. In particular, we have the following proof of the two-square theorem via factorization in $\mathbb{Z}[i]$:

Let p be a prime congruent to 1 mod 4. As we have seen in our earlier proof of the two-square theorem, there exists n such that p divides n^2+1. Thus, in $\mathbb{Z}[i]$, we have that p divides $(n+i)(n-i)$. Since neither $\frac{n+i}{p}$ nor $\frac{n-i}{p}$ lie in $\mathbb{Z}[i]$, p does not divide $n+i$ or $n-i$ in $\mathbb{Z}[i]$. Thus p is not prime in $\mathbb{Z}[i]$. By the corollary p must be a sum of two squares.

This proof looks non-constructive, but in fact it isn't; if you want to find a, b such that $p = a^2+b^2$, simply use Euclid's algorithms to find a GCD of p and $n+i$. Since p divides neither $n+i$ nor $n-i$, this GCD is not a unit, nor is it an associate of p, so its norm is neither 1 nor p^2; therefore this GCD has norm exactly p. In fact, this is precisely what the Fermat descent in our original proof is doing!

Since we have also shown that primes congruent to 3 mod 4 are not the sum of two squares, we have the following complete characterization of sums of two squares:

Theorem 10.3. *An integer n is a sum of two squares if and only if its prime factorization is of the form:*

$$n = 2^r p_1^{s_1} \dots p_k^{s_k} q_1^{t_1} \dots q_h^{t_h}$$

where the p_i are primes congruent to 1 mod 4, the q_i are primes congruent to 3 mod 4, and the t_i are even.

Proof. Suppose n is of the above form, and write:

$$z = (1+i)^r (a_1+b_1 i)^{s_1} \dots (a_k+b_k i)^{s_k} q_1^{\frac{t_1}{2}} \dots q_h^{\frac{t_h}{2}},$$

where $p_i = N(a_i+b_i)$. Then $N(z) = n$. Conversely, if n is a sum of two squares, we can write $n = N(z)$, and factor $z = u(1+i)^r p_1^{s_1} \dots p_k^{s_k} q_1^{t_1} \dots q_h^{t_h}$ in $\mathbb{Z}[i]$, where the p_i are primes in $\mathbb{Z}[i]$ whose norm is a prime congruent to 1 mod 4, and the q_i are primes in $\mathbb{Z}$ congruent to 3 mod 4. Then $N(z)$ has the claimed form. □

10.2. **Representing primes by quadratic forms.** The advantage of this newer proof of the two square theorem, and the related characterization of sums of two squares is that it is now clear how to generalize this approach to other Euclidean domains. Let α be an imaginary quadratic algebraic integer, and a, b integers such that $\alpha^2 + a\alpha + b = 0$. We have

$$N(x-y\alpha) = (x-y\alpha)(x-y\overline{\alpha}) = x^2 + axy + by^2;$$

we will refer to the polynomial $P(x,y) = x^2+axy+by^2$ as the *norm form* for α.

Theorem 10.4. *Suppose that unique factorization holds in $\mathbb{Z}[\alpha]$, and let p be an integer prime such that the polynomial $P(x,y) = x^2+ax+b$ has a root mod p. Then there exist integers x and y such that $P(x,y) = p$.*

Proof. We have $x^2 + ax + b = (x - \alpha)(x - \overline{\alpha})$. So if $x^2 + ax + b$ has a root mod p, then there exists x such that p divides $(x - \alpha)(x - \overline{\alpha})$ in $\mathbb{Z}[\alpha]$. Since neither $\frac{x-\alpha}{p}$ nor $\frac{x-\alpha}{p}$ lies in $\mathbb{Z}[\alpha]$, p is not prime in $\mathbb{Z}[\alpha]$. Therefore p must be reducible in $\mathbb{Z}[\alpha]$, so $p = uv$ with neither u nor v units. Then $N(u)N(v) = p^2$, and neither $N(u)$ nor $N(v) = 1$, so $N(u) = p$ and the result follows. □

For instance, if $\alpha = \frac{-1+\sqrt{-3}}{2}$, then $a = b = 1$ and $\mathbb{Z}[\alpha]$ is a Euclidean domain (example sheet 3 Q6). We thus see that any prime such that x^2-x+1 has a root mod p is of the form $x^2 + xy + y^2$. On the other hand, we have:

Lemma 10.5. *Let p be an odd prime. The polynomial $x^2 - ax + b$ has a root mod p if, and only if, $a^2 - 4b$ is a quadratic residue mod p.*

Proof. Suppose $a^2 - 4b \equiv d^2 \pmod{p}$, and let $x = 2^{-1}(a + d)$. Then $x^2 - ax + b = 4^{-1}(a^2 + 2ad + d^2 - 2a(a + d) + 4b) = 0$. Conversely, if x is a root of $x^2 - ax + b$, then $(2x - a)^2 = 4x^2 - 4ax + a^2 = a^2 - 4b$. □

Thus in particular x^2+x+1 has a root mod p if, and only if, -3 is a square mod p; by quadratic reciprocity this holds precisely when $p = 3$ or $p \equiv 1 \pmod{3}$. Conversely, by directly checking mod 3 we see that $x^2 + xy + y^2$ is always 0 or 1 mod 3, so the primes of the form $x^2 + xy + y^2$ are precisely 3 and the primes congruent to 1 mod 3.

When unique factorization fails in $\mathbb{Z}[\alpha]$ the above result is false. For instance, when $\alpha = \sqrt{-5}$ ($a = 0$, $b = 5$), then $P(x) = x^2 + 5y^2$. In particular 3 is not of the form $P(x, y)$, even though $x^2 + 5$ has a root mod 3. In this situation the question of finding the primes of the form $P(x, y)$ has connections to class field theory and is a very deep part of modern algebraic number theory. For a taste of this, Cox's book *Primes of the form $x^2 + ny^2$* is very good.

11. Pell's equation

11.1. **Pell's Equation.** Let $d > 1$ be a squarefree integer, and consider the equation $x^2 - dy^2 = 1$. This is called Pell's equation.

Note that $\mathbb{Z}[\sqrt{d}]$ is a real quadratic subring of $\mathbb{C}$, and its norm form is given by $N(x + y\sqrt{d}) = x^2 - dy^2$. Thus the problem of finding integer solutions to Pell's equation is equivalent to finding elements of norm 1 in $\mathbb{Z}[\sqrt{d}]$. Since such elements are units, and the norm is multiplicative, these elements form a subgroup $\mathbb{Z}[\sqrt{d}]^{\times,1}$ of $\mathbb{Z}[\sqrt{d}]^{\times}$.

There are two obvious elements of $\mathbb{Z}[\sqrt{d}]^{\times,1}$, namely ± 1. All others are of the form $x + y\sqrt{d}$, with x and y integers and y nonzero. We have:

Lemma 11.1. *Let $x + y\sqrt{d}$ be an element of $\mathbb{Z}[\sqrt{d}]^{\times,1}$. Then:*

- *We have $x, y > 0$ if and only if $x + y\sqrt{d} > 1$.*
- *We have $x > 0, y < 0$ if and only if $0 < x + y\sqrt{d} < 1$.*
- *We have $x < 0, y > 0$ if and only if $-1 < x + y\sqrt{d} < 0$.*

- *We have $x, y < 0$ if and only if $x + y\sqrt{d} < -1$.*

Proof. It is clear that if $x, y > 0$ then $x + y\sqrt{d} > 1$. But then $x - y\sqrt{d} = (x+y\sqrt{d})^{-1}$ lies between 0 and -1. Similary, $-x+y\sqrt{d}$ lies between -1 and 0, and $-x - y\sqrt{d}$ is less than -1. Thus we have the rightward implication in each of the four claims above. But since the four cases are mutually exclusive and exhaust all the possibilities, the leftward implications hold as well. □

Lemma 11.2. *Let $z = x+y\sqrt{d}$, $z' = x'+y'\sqrt{d}$ be two elements of $\mathbb{Z}[\sqrt{d}]^{\times,1}$ with x, y, x', y' all positive. Then $z > z'$ if, and only if, $y > y'$.*

Proof. We have $z - \frac{1}{z} = x + y\sqrt{d} - (x - y\sqrt{d}) = 2y\sqrt{d}$. Since $z - \frac{1}{z}$ is an increasing function for z positive, we have $z > z'$ iff $z - \frac{1}{z} > z' - \frac{1}{z}'$ iff $y > y'$. □

Suppose we have a nontrivial element of $\mathbb{Z}[\sqrt{d}]^{\times,1}$ (that is, one other than ± 1). Without loss of generality we can take x and y positive. There then exists $\epsilon = x + y\sqrt{d} \in \mathbb{Z}[\sqrt{d}]^{\times,1}$ such that x and y are both positive and y is as small as possible. We will call ϵ the *fundamental 1-unit* in $\mathbb{Z}[\sqrt{d}]$. By the previous two lemmas it is the smallest element of $\mathbb{Z}[\sqrt{d}]^{\times,1}$ that is greater than one. For example in $\mathbb{Z}[\sqrt{2}]$ we have $\epsilon = 3 + 2\sqrt{2}$.

Proposition 11.3. *Suppose there exists a nontrivial element of $\mathbb{Z}[\sqrt{d}]^{\times,1}$. Then every element of $\mathbb{Z}[\sqrt{d}]^{\times,1}$ is of the form $\pm\epsilon^n$ for some n in $\mathbb{Z}$, where ϵ is the fundamental 1-unit.*

Proof. Let z be an element of $\mathbb{Z}[\sqrt{d}]^{\times,1}$. After negating z and/or replacing z by $\frac{1}{z}$, as necessary, we can assume $z > 1$. Since ϵ is also greater than one, there exists n such that $\epsilon^n \leq z < \epsilon^{n+1}$. Then $\epsilon^{-n}z$ is a 1-unit with $1 \leq \epsilon^{-n}z < \epsilon$; since ϵ is the smallest 1-unit greater than one we have $\epsilon^{-n}z = 1$. □

11.2. **Constructing the fundamental 1-unit.** In fact, we will show that there are always elements of $\mathbb{Z}[\sqrt{d}]^{\times,1}$ other than ± 1, so that we are always in the situation of the preceding proposition. The key idea is to note that $x^2 - dy^2 = 0$ precisely when $\frac{x}{y}$ is a square root of d, and thus, when d is squarefree, $x^2 - dy^2 = 1$ when $\frac{x}{y}$ is in some sense "as close as possible" to $\sqrt{d}$. This suggests we should think about approximating $\sqrt{d}$ by rational numbers.

It's clear that if p and q are integers, and α is an irrational number, then we can make $|\frac{p}{q} - \alpha|$ as small as we want. However, in order to do so we might need to make q large. We will thus be interested in approximations to α where the error $|\frac{p}{q} - \alpha|$ is small *compared to* $\frac{1}{q^n}$ *for various* n. As n gets larger it will become harder and harder to find such approximations.

When $n = 1$ the situation is very easy: for any α, and any q, there exists p such that $|\frac{p}{q} - \alpha| < \frac{1}{q}$.

When $n = 2$ things are much less trivial, but we have the following important result, due to Dirichlet:

Theorem 11.4. *Let α be an irrational number, and $Q > 1$ an integer. Then there exist p, q integers, with $1 \leq q < Q$, such that $|p - q\alpha| < \frac{1}{Q}$.*

Proof. For $1 \leq k \leq Q-1$, let $a_k = \lfloor k\alpha \rfloor$, so that $0 < k\alpha - a_k < 1$.

Partition the interval $[0,1]$ into Q subintervals of length $\frac{1}{Q}$. One of these intervals contains two elements of the set:

$$\{0, \alpha - a_1, 2\alpha - a_2, \ldots, (Q-1)\alpha - a_{Q-1}, 1\}$$

The difference between these two elements is of the form $p - q\alpha$, where p and q are integers and q is less than Q, and this difference is less than $\frac{1}{Q}$. □

Corollary 11.5. *For any irrational α there are infinitely many $\frac{p}{q}$ such that $|\alpha - \frac{p}{q}| < \frac{1}{q^2}$.*

Proof. It suffices to show, given any $\frac{p}{q}$ with $|\alpha - \frac{p}{q}| < \frac{1}{q^2}$, that we can find another $\frac{p'}{q'}$ with $|\alpha - \frac{p'}{q'}| < \frac{1}{(q')^2} < |\alpha - \frac{p}{q}|$.

Suppose given such a $\frac{p}{q}$, and choose Q such that $\frac{1}{Q} < |\alpha - \frac{p}{q}|$. Then by the theorem there exist p', q' with $q' < Q$, and $|\alpha - \frac{p'}{q'}| < \frac{1}{Qq'} < \frac{1}{(q')^2}$. Also $|\alpha - \frac{p'}{q'}| < \frac{1}{Qq'} < \frac{1}{Q} < |\alpha - \frac{p}{q}|$. The claim follows. □

We can now show:

Theorem 11.6. *For any squarefree d there is a nontrivial solution to $x^2 - dy^2 = 1$.*

Proof. The corollary gives us infinitely many pairs (p_i, q_i) such that $|p_i - q_i\sqrt{d}| < \frac{1}{q_i}$. Note that then $|p_i + q_i\sqrt{d}| < \frac{1}{q_i} + 2q_i\sqrt{d} < 3q_i\sqrt{d}$. We thus have $|N(p_i - q_i\sqrt{d})| = |(p_i - q_i\sqrt{d})(p_i + q_i\sqrt{d})| < 3\sqrt{d}$.

Thus for some M between $-3\sqrt{d}$ and $3\sqrt{d}$ there are infinitely many pairs (p_i, q_i) such that $N(p_i + q_i\sqrt{d}) = M$.

Since there are finitely many congruence classes mod M, there is some pair (p_0, q_0) such that there are infinitely many pairs (p_i, q_i) with $N(p_i + q_i\sqrt{d}) = M$, $p_i \equiv p_0 \pmod{M}$, and $q_i \equiv q_0 \pmod{M}$.

Now for (p_i, q_i) and (p_j, q_j) any two such pairs, consider the quotient:

$$\frac{p_i - q_i\sqrt{d}}{p_j - q_j\sqrt{d}} = \frac{(p_ip_j - dq_iq_j) + (p_iq_j - p_jq_i)\sqrt{d}}{M}.$$

The congruence conditions show that this quotient lies in $\mathbb{Z}[\sqrt{d}]$, and it has norm 1 by multiplicativity of the norm. □

11.3. **The equation $x^2 - dy^2 = -1$.** Note that we can also apply these techniques to solving $x^2 - dy^2 = -1$. Solutions correspond to elements of norm -1 in $\mathbb{Z}[\sqrt{d}]$; if $x + y\sqrt{d}$ is one solution, then since this is a unit, the others are given by $\pm(x + y\sqrt{d})\epsilon^n$, where ϵ is the fundamental 1-unit. Note, however, that unlike for 1 units there may be no -1-units at all (consider, for instance, $d = 3$, where the equation has no solutions mod 3 and thus no integer solutions; but when $d = 2$ we can take $x = y = 1$.)

12. Continued Fractions

12.1. **Rational Continued Fractions.** Given a rational number $\frac{p}{q}$, we can write $\frac{p}{q} = a_0 + r_0$, where a_0 is an integer and $0 \leq r_0 < 1$ is between 0 and 1. If r_0 is not zero, then we can write $\frac{1}{r_0} = a_1 + r_1$, with a_1 and integer and $0 \leq r_1 < 1$. We then have:

$$\frac{p}{q} = a_0 + r_0 = a_0 + \frac{1}{a_1 + r_1}.$$

Continuing in this way, as long as r_i is nonzero we set $\frac{1}{r_i} = a_{i+1} + r_{i+1}$, with a_i an integer and $0 \leq r_{i+1} < 1$. The denominators of the r_i are strictly decreasing, so eventually some $r_n = 0$ and we have:

$$\frac{p}{q} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}}.$$

This expression is called the *continued fraction expansion* of $\frac{p}{q}$. It is closely related to Euclid's algorithm; indeed, it's not hard to see that the a_i are the quotients q_i from Euclid's algorithm applied to the pair p, q. Note that each a_i is an integer and for $i \geq 1$, $a_i \geq 1$. Example: $40/19 = 2 + 1/(19/2) = 2 + \frac{1}{9 + \frac{1}{2+0}}$.

12.2. **Infinite Continued Fractions.** Now let α be an irrational real number. As above, we can write $\alpha = a_0 + r_0$, where a_0 is an integer and $0 \leq r_0 < 1$, and then for each i set $\frac{1}{r_i} = a_{i+1} + r_{i+1}$ with a_{i+1} an integer and $0 \leq r_{i+1} < 1$. Note that unlike in the rational case, this sequence will never terminate, as r_i is always irrational and thus never zero. We write:

$$\alpha = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \ldots}}$$

and call this the *continued fraction expansion* of α. Note that so far this is just a formal expression! In fact, we can make mathematical sense of this expression, but it requires some justification.

Example: $\sqrt{3}$. $\sqrt{3} = 1 + (\sqrt{3} - 1)$, $1/(\sqrt{3} - 1) = (\sqrt{3} + 1)/2 = 1 + (\sqrt{3} - 1)/2$, $2/(\sqrt{3} - 1) = \sqrt{3} + 1 = 2 + (\sqrt{3} - 1)$, and we carry on getting alternating 1, 2 forever.

First, we introduce some useful notation: For $a_0, \dots a_n$ real numbers, we define $[a_0; a_1, \dots, a_n]$ to be the real number:

$$a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}}$$

when this expression is well-defined (that is, when it does not involve a division by zero.)

The value of this expression can computed by a recurrence relation, as follows:

Lemma 12.1. *Given $a_0, \dots, a_n$ real numbers, define p_i, q_i for $0 \leq i \leq n$ by $p_0 = a_0$, $q_0 = 1$, $p_1 = a_0 a_1 + 1$, $q_1 = a_1$, together with the recurrences:*

$$p_i = a_i p_{i-1} + p_{i-2}$$

$$q_i = a_i q_{i-1} + q_{i-2}.$$

Then $[a_0; a_1, \dots, a_n] = \frac{p_n}{q_n}$, assuming no q_i is zero.

Proof. We prove this by induction on n; the cases $n = 0$ and $n = 1$ are clear. Let p_i' and q_i' be the numbers defined by the recurrence attached to the sequence $a_0, a_1, \dots, a_{n-2}, a_{n-1} + \frac{1}{a_n}$. The induction hypothesis tells us that $[a_0; a_1, \dots, a_{n-2}, a_{n-1} + \frac{1}{a_n}] = \frac{p'_{n-1}}{q'_{n-1}}$. By the recurrence defining the p' and q', we have

$$\frac{p'_{n-1}}{q'_{n-1}} = \frac{(a_{n-1} + \frac{1}{a_n})p'_{n-2} + p'_{n-3}}{(a_{n-1} + \frac{1}{a_n})q'_{n-2} + q'_{n-3}}$$

Since the sequences defining p_i',q_i' and p_i,q_i agree for $i \leq n-2$, the latter is equal to:

$$\frac{(a_{n-1} + \frac{1}{a_n})p_{n-2} + p_{n-3}}{(a_{n-1} + \frac{1}{a_n})q_{n-2} + q_{n-3}}.$$

Multiplying both numerator and denominator by a_n, and using $p_{n-1} = a_{n-1}p_{n-2} + p_{n-3}$, $p_n = a_n p_{n-1} + p_{n-2}$ (and similarly for q_n) we find that this expression is equal to $\frac{p_n}{q_n}$.

Thus we have:

$$[a_0; a_1, \dots, a_n] = [a_0; a_1, \dots, a_{n-2}, a_{n-1} + \frac{1}{a_n}] = \frac{p_n}{q_n}$$

as claimed. □

Note that if $a_i \geq 1$ for all i (as will be the case, for instance, if the a_i come from taking a continued fraction expansion of some real number), then the q_i are all nonzero and form a strictly increasing sequence. (Indeed, if all a_i are at least one, then $q_i \geq q_{i-1} + q_{i-2} \geq 2q_{i-2}$, so this sequence increases exponentially fast.)

Now suppose we have an infinite sequence $a_0, a_1, a_2, \dots$, of real numbers, and assume that $a_i \geq 1$ for $i \geq 1$. Define p_i and q_i by the recurrence given above. We call $\frac{p_i}{q_i}$ the *ith convergent* to the continued fraction:

$$a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots}}.$$

Lemma 12.2. *We have* $p_n q_{n-1} - q_n p_{n-1} = (-1)^{n-1}$.

Proof. This is again by induction on n; the base case $n = 1$ is clear. Assume this is true for $n - 1$. Then by the defining recurrence for the p_i and q_i, we have:

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

and the claim follows. □

(Note that in particular this implies that $(p_n, q_n) = 1$.) It follows that $|\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}| < \frac{1}{q_n q_{n-1}}$. Since the q_n increase exponentially, it follows that the $\frac{p_i}{q_i}$ form a Cauchy sequence, and hence converge to a real number.

A natural question to ask at this point is: "if the sequence $a_0, a_1, \dots$ arises from the continued fraction expansion of an irrational number α, is the limit of the convergents $\frac{p_n}{q_n}$ equal to α?". This is indeed the case:

Lemma 12.3. *Let α be an irrational real number, and let $a_0, a_1, \dots$ be the sequence of integers arising from its continued fraction expansion. Let $\frac{p_n}{q_n}$ be the nth convergent. Then $\frac{p_n}{q_n} < \alpha$ if n is even, and $\frac{p_n}{q_n} > \alpha$ if n is odd.*

Proof. Once again we use induction on n; the case $n = 0$ is clear. Note that $[a_1; a_2, \dots, a_n]$ is the $(n-1)$st convergent to $\frac{1}{\alpha - a_0}$. Thus, by the induction hypothesis, if n is odd we have:

$$[a_1; a_2, \dots, a_n] < \frac{1}{\alpha - a_0}$$

and thus

$$\alpha < a_0 + \frac{1}{[a_1; a_2 \dots, a_n]} = [a_0; a_1, a_2, \dots, a_n].$$

If n is even the same argument works with the inequalities reversed. □

Corollary 12.4. *Let α be an irrational real number, and let $a_0, a_1, \dots$ be the sequence of integers arising from its continued fraction expansion. Let $\frac{p_n}{q_n}$ be the nth convergent. Then $|\alpha - \frac{p_n}{q_n}| < \frac{1}{q_n q_{n+1}}$. In particular the limit $\frac{p_n}{q_n}$ as n approaches infinity is α.*

Proof. Since exactly one of $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$ is less than α, we have

$$|\alpha - \frac{p_n}{q_n}| < |\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}| = \frac{1}{q_n q_{n+1}}.$$

Since the q_n increase exponentially quickly the second claim is clear. □

Note that since $\frac{1}{q_n q_{n+1}} < \frac{1}{q_n^2}$, this gives a second, *completely constructive* proof of Dirichlet's theorem on rational approximations!

12.3. **Best Approximations.** In fact, there is a sense in which the convergents to the continued fraction expansion of α are the best rational approximations to α. We will make this precise below.

Fix α real and irrational, and as above define a_i and r_i by setting $\alpha = a_0 + r_0$, with a_0 an integer and $0 \leq r_0 < 1$, and $\frac{1}{r_i} = a_{i+1} + r_{i+1}$ with a_{i+1} an integer and $0 \leq r_{i+1} < 1$. Thus $\alpha = [a_0; a_1, \ldots, a_n, 1/r_n]$. Define p_n and q_n from the sequence $a_0, a_1, \ldots$ by the recurrences of the previous section.

Lemma 12.5. *For all n, we have:*

$$\alpha = \frac{p_n + p_{n-1} r_n}{q_n + q_{n-1} r_n}.$$

Proof. This follows from Lemma 12.1 applied to $\alpha = [a_0; a_1, \ldots, a_n, 1/r_n]$. □

Corollary 12.6. *For all n, $|\alpha q_n - p_n| < |\alpha q_{n-1} - p_{n-1}|$, so that $|\alpha - \frac{p_n}{q_n}| < |\alpha - \frac{p_{n-1}}{q_{n-1}}|$.*

Proof. By Lemma 12.5, we have $|\alpha q_n - p_n| = r_n |\alpha q_{n-1} - p_{n-1}|$, and $r_n < 1$. Since $q_n > q_{n-1}$ the second inequality follows. □

Theorem 12.7. *Let h, k be integers with $0 < |k| < q_{n+1}$. Then $|k\alpha - h| \geq |\alpha q_n - p_n|$ with equality only if $|k| = q_n$.*

If $|k| \leq q_n$ then $|\frac{h}{k} - \alpha| \geq |\frac{p_n}{q_n} - \alpha|$ with equality if and only if $\frac{h}{k} = \frac{p_n}{q_n}$; in other words, $\frac{p_n}{q_n}$ is the best rational approximation to α with denominator at most q_n.

Proof. By Lemma 12.2 we can find integers u, v with

$$h = u p_n + v p_{n+1}$$

$$k = u q_n + v q_{n+1}$$

(because the corresponding matrix for these linear equations has determinant $(-1)^{n-1}$). Since $|k| < q_{n+1}$, and there is nothing to prove if $k = -$, we have $u \neq 0$, and also u, v must have opposite signs if $v \neq 0$. If $v = 0$ then we are done. Otherwise, since $\alpha q_n - p_n$ and $\alpha q_{n+1} - p_{n+1}$ have opposite signs by Lemma 12.3, we see that $u(\alpha q_n - p_n)$ and $v(\alpha q_{n+1} - p_{n+1})$ have the same sign. Since

$$u(\alpha q_n - p_n) + v(\alpha q_{n+1} - p_{n+1}) = k\alpha - h,$$

the claim follows (with equality requiring $v = 0$).

If furthermore $|k| \leq q_n$ then we can multiply this inequality and the inequality $1/|k| \geq 1/q_n$ to get the claim. □

Corollary 12.8. *If h, k are integers with $|\alpha - h/k| < \frac{1}{2k^2}$, then $h/k = p_n/q_n$ for some n.*

Proof. WLOG $k \geq 1$, and we can choose n with $q_n \leq k < q_{n+1}$. Then we have

$$|p_n/q_n - h/k| \leq |p_n/q_n - \alpha| + |\alpha - h/k| = 1/q_n|p_n - q_n\alpha| + 1/k|h - k\alpha| \leq (1/q_n + 1/k)|h - k\alpha|$$

by Theorem 12.7, so our hypothesis gives $|p_n/q_n - h/k| < (1/q_n + 1/k)(1/2k) \leq 1/kq_n$. But this forces $p_n/q_n = h/k$, as required. □

12.4. **Returning to Pell's equation.** We can use Corollary 12.8 to give us a better algorithm for finding solutions to Pell's equation, and as a bonus it lets us solve $x^2 - dy^2 = -1$ too (when that has solutions).

Proposition 12.9. *Let $d > 1$ be squarefree, and let p_n/q_n be the convergents to $\sqrt{d}$. If $x, y > 0$ and $x^2 - dy^2 = \pm 1$, then $x = p_n$, $y = q_n$ for some n.*

Proof. Note that it is enough to show that $x/y = p_n/q_n$, as then $x = rp_n$, $y = rq_n$ for some $r \geq 1$, and $x^2 - dy^2 = r^2(p_n^2 - dq_n^2)$ is divisible by r^2, so $r = 1$.

Suppose firstly that $x^2 - dy^2 = 1$. Then $x - d\sqrt{y} = 1/(x + y\sqrt{d}) > 0$, so $x > d\sqrt{y}$, and we have

$$|x/y - \sqrt{d}| = 1/y(x + y\sqrt{d}) < 1/2y^2\sqrt{d} < 1/2y^2,$$

and the claim follows from Corollary 12.8.

Now suppose that $x^2 - dy^2 = -1$. Here we will have to use a trick: rewrite the equation as $y^2 - (1/d)x^2 = 1/d$. Then $y > \frac{1}{\sqrt{d}}x$, and so

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

so that Corollary 12.8 implies that y/x is a convergent to the continued fraction of $1/\sqrt{d}$. But $0 < 1/\sqrt{d} < 1$, so we see that if $[a_0; a_1, \dots]$ is the continued fraction $\sqrt{d}$, then the continued fraction for $1/\sqrt{d}$ is $[0; a_0, a_1, \dots]$, so that its convergents are just the q_n/p_n. So $y/x = q_n/p_n$, as required. □

So we now have an algorithm: we just have to go through the convergents, and see if they satisfy $x^2 - dy^2 = \pm 1$. However, we don't yet know just how far along we have to go. In fact, it turns out that there the continued fractions for $\sqrt{d}$ have a particularly nice form. We already saw that $\sqrt{3} = [1; 1, 2, 1, 2, 1, 2, \dots]$; let's write this as $\sqrt{3} = [1; \overline{1, 2}]$. Let's do some more examples.

$\sqrt{2} = 1 + (\sqrt{2} - 1)$, $1/(\sqrt{2} - 1) = \sqrt{2} + 1 = 2 + (\sqrt{2} - 1)$, so we see that $\sqrt{2} = [1; \overline{2}]$. Similarly $\sqrt{5} = 2 + (\sqrt{5} - 2)$, and $1/(\sqrt{5} - 2) = \sqrt{5} = 4 + (\sqrt{5} - 2)$, so $\sqrt{5} = [2, \overline{4}]$.

Here are a couple more examples: it turns out that $\sqrt{7} = [2; \overline{1, 1, 1, 4}]$, $\sqrt{13} = [3; \overline{1, 1, 1, 1, 6}]$, and $\sqrt{43} = [6; \overline{1, 1, 3, 1, 5, 1, 3, 1, 1, 12}]$

We can start to see a pattern, or in fact several patterns. The following could be proved in a couple more lectures using similar techniques to the ones we've been using.

Definition 12.10. The continued fraction expansion $[a_0; a_1, a_2, \dots]$ of an irrational number α is *eventually periodic* if there exist positive integers N and d such that $a_n = a_{n+d}$ for all $n \geq N$. It is *periodic* if there exists a positive integer d such that $a_n = a_{n+d}$ for all n.

Fact 12.11. (1) *The continued fraction of $\sqrt{d}$ is eventually periodic.*
(2) *In fact, it is of the form $[a_0; \overline{a_1, \dots, a_{m-1}, 2a_0}]$.*
(3) *The sequence $a_1, \dots, a_{m-1}$ is symmetric, i.e. $a_i = a_{m-i}$.*
(4) *The n for which $p_n^2 - dq_n^2 = \pm 1$ are precisely the $n \equiv -1 \pmod{m}$, in which case $p_{lm-1}^2 - dq_{lm-1}^2 = (-1)^{lm}$. In particular,*
(5) *the fundamental 1-unit is given by $p_{m-1} + q_{m-1}\sqrt{d}$ if m is even and $p_{2m-1} + q_{2m-1}\sqrt{d}$ if m is odd.*
(6) *There is a solution to $x^2 - dy^2 = -1$ if and only if m is odd, in which case the n for which $p_n^2 - dq_n^2 = 1$ are precisely the $n \equiv m - 1 \pmod{2m}$.*

Let's see how to use this to find solutions for the example of $x^2 - 43y^2 = \pm 1$. Since $\sqrt{43} = [6; \overline{1, 1, 3, 1, 5, 1, 3, 1, 1, 12}]$ we have $m = 10$, which is even, so we will not be a solution to $x^2 - 43y^2 = -1$. We can use the usual recurrence of Lemma 12.1. The q_i are smaller, so we may as well start with them. $q_0 = 1$, $q_1 = a_1 = 1$, $q_2 = a_2q_1 + q_0 = 2$, $q_3 = a_3q_2 + q_1 = 7$, $q_4 = a_4q_3 + q_2 = 9$, $q_5 = 52$, $q_6 = 61$, $q_7 = 235$, $q_8 = 296$, $q_9 = 531$. It turns out that $p_9 = 3482$, and indeed $3482^2 - 43 \cdot 531^2 = 1$.

Similarly we can do $x^2 - 13y^2 = \pm 1$. This time $\sqrt{13} = [3; \overline{1, 1, 1, 1, 6}]$ and $m = 5$ so there is a solution to $x^2 - 13y^2 = -1$, and it will come from p_4, q_4. In this case the recurrence is really easy, and in fact $q_1 = 1$, $q_2 = 2$,$q_3 = 3$ $q_4 = 5$, and $p_4 = 18$, and $18^2 - 13 \cdot 5^2 = -1$. Similarly $q_5 = 33$, $q_6 = 38$, $q_7 = 71$, $q_8 = 109$, $q_9 = 180$, while $p_9 = 649$, and $649^2 - 13 \cdot 180^2 = 1$. Note also that as expected, $649 + 180\sqrt{13} = (18 + 5\sqrt{13})^2$, and this is a much faster way of calculating q_9!

12.5. **Periodic Continued Fractions.** While we aren't going to prove that $\sqrt{d}$ has an eventually periodic continued fraction, we do prove the following easier converse statement.

Definition 12.12. An irrational number α is a *quadratic irrational* if there is a polynomial $ax^2 + bx + c$, with a, b, c integers, that has α as a root.

Proposition 12.13. *Suppose that α has an eventually periodic continued fraction expansion. Then α is a quadratic irrational.*

Proof. We first show this when α is periodic. We then have a d such that

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{a_{d-1} + \cfrac{1}{\alpha}}}}.$$

Simplifying the right hand side (or just applying Lemma 12.5), we find integers x, y, z, w such that

$$\alpha = \frac{x\alpha + y}{z\alpha + w}.$$

Then $z\alpha^2 + (w-x)\alpha - y = 0$; since α is irrational z cannot be zero, so α is a quadratic irrational.

If α is only eventually periodic there is a β with periodic continued fraction expansion such that we have:

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{a_{N-1} + \cfrac{1}{\beta}}}}.$$

Then β is quadratic irrational. It thus suffices to show that if β is quadratic irrational, so is $\frac{1}{\beta}$ and $n + \beta$ for integer n.

Note that if β is a root of $ax^2 + bx + c$, then $\frac{1}{\beta}$ is a root of $a + bx + cx^2$, and $n + \beta$ is a root of $a(x-n)^2 + b(x-n) + c$, so these claims are clear. □

In fact, all quadratic irrationals have eventually periodic continued fraction expansions, but we will not prove this.

13. Diophantine Approximation

13.1. **Liouville's theorem.** We've shown (in two different ways!) that for any irrational real number α, there are infinitely many rational numbers $\frac{p}{q}$ with $|\frac{p}{q} - \alpha| < \frac{1}{q^2}$. What happens if we ask for something stronger? For instance, can we find infinitely many $\frac{p}{q}$ with $|\frac{p}{q} - \alpha| < \frac{1}{q^e}$ for some $e > 2$?

It turns out that there for many irrational α, the answer is *no*. In particular, let's make the following definition:

Definition 13.1. Let d be a positive integer. A complex number α is *algebraic of degree d* if there is a degree d (not necessarily monic) polynomial $P(x)$, with integer coefficients, such that $P(\alpha) = 0$, and no such polynomial of degree less than d.

We then have:

Theorem 13.2 (Liouville's theorem)**.** *Let α be a real number that is algebraic of degree d. Then for any real number $e > d$, there are at most finitely many rational numbers $\frac{p}{q}$ such that $|\frac{p}{q} - \alpha| < \frac{1}{q^e}$.*

Proof. Let $P(x)$ be a polynomial of degree d, with integer coefficients, such that $P(\alpha) = 0$. Choose ϵ such that $P(x)$ has no roots other than α on the closed interval $[\alpha - \epsilon, \alpha + \epsilon]$.

Write $P(x) = (x - \alpha)Q(x)$; $Q(x)$ is a monic polynomial with real coefficients, of degree $d - 1$. Since $|Q(x)|$ is a continuous, real valued function, there is a real number $K > 0$ such that $|Q(x)| \leq K$ on the compact set $[\alpha - \epsilon, \alpha + \epsilon]$.

Now suppose we have $\frac{p}{q}$ with $|\frac{p}{q} - \alpha| < \frac{1}{q^e}$. There are only finitely many q such that $\frac{1}{q^e} \geq \epsilon$, so we may assume $\frac{1}{q^e} < \epsilon$.

Now on the one hand, we have:

$$|P(\frac{p}{q})| = |(\frac{p}{q} - \alpha)Q(\frac{p}{q})| < \frac{1}{q^e}K.$$

On the other hand, since P has degree d and integer coefficients, the denominator of $P(\frac{p}{q})$ (when written in lowest terms) is a divisor of q^d. But $\frac{p}{q}$ is NOT a root of $P(x)$, so $P(\frac{p}{q})$ is nonzero and hence $|P(\frac{p}{q})| \geq \frac{1}{q^d}$.

Putting the inequalities together, we get:

$$\frac{1}{q^d} \leq |P(\frac{p}{q})| < \frac{1}{q^e}K.$$

Rewriting, we find $q^{e-d} < K$; since $e - d > 0$ there are only finitely many q for which this is possible. □

13.2. **Constructing transcendentals.** Recall that a complex number α is *transcendental* if there is no polynomial $P(x)$, with integer coefficients, such that $P(\alpha) = 0$, and *algebraic* otherwise.

The set of polynomials $P(x)$ with integer coefficients is countable; since each such polynomial has finitely many roots the set of algebraic numbers is countable. Since the set of reals is uncountable this means that, in a very strong sense, almost every real number is transcendental.

In spite of this it is very hard to give an example of a single real number that is provably transcendental. (In fact, e and π are examples of transcendental numbers, but this is much harder than what we'll do.)

Liouville's theorem gives one approach to proving that a given number is transcendental: show that it admits too many good rational approximations. If we can show that for any e, there exist infinitely many $\frac{p}{q}$ such that $|\frac{p}{q} - \alpha| < \frac{1}{q^e}$, then Liouville's theorem tells us that α can't be algebraic of any degree, and hence must be transcendental.

As an example, define a real number α by

$$\alpha = \sum_{n=1}^{\infty} \frac{1}{10^{n!}}.$$

This clearly converges.

We can find rational approximations to α simply by truncating the series; for each k, let α_k be the sum:

$$\alpha_k = \sum_{n=1}^{k} \frac{1}{10^{n!}}.$$

On one hand, α_k is rational with denominator $10^{k!}$. On the other hand, we have:

$$|\alpha - \alpha_k| = \sum_{n=k+1}^{\infty} \frac{1}{10^{n!}} < \frac{2}{10^{(k+1)!}}.$$

Fix a positive integer d. For any $k > d$, we have $\frac{2}{10^{(k+1)!}} < \frac{1}{(10^{k!})^d}$. Thus there are infinitely many k such that $|\alpha - \alpha_k| < \frac{1}{(10^{k!})^d}$. Liouville's theorem thus tells us that α cannot be algebraic of degree d. Since d was arbitrary, α must be transcendental.

13.3. **Roth's theorem.** Liouville's theorem tells us that algebraic numbers are difficult to approximate well by rationals. In fact, they are even more difficult to approximate than Liouville's theorem would suggest. For instance, we have:

Theorem 13.3 (Roth's Theorem)**.** *Suppose α is algebraic. Then for any $\epsilon > 0$, there are only finitely many rational numbers $\frac{p}{q}$ such that $|\frac{p}{q} - \alpha| < \frac{1}{q^{2+\epsilon}}$.*

In other words, for algebraic numbers, you cannot do any better than Dirichlet's theorem.

Roth's theorem is considerably harder to prove than Liouville's, and we will not give a proof in this course. Note that the stronger bound gives proofs that additional real numbers are transcendental; for instance, one can prove with Roth's theorem that the number

$$\beta = \sum_{n=1}^{\infty} \frac{1}{10^{3^n}}$$

is transcendental, which is not possible with Liouville's theorem.

This also shows that higher degree versions of Pell's equation can only have finitely many solutions. For example, suppose that $d > 0$ is not a cube; then there are only finitely many pairs (x, y) of integers with $x^3 - dy^3 = 1$. Indeed we can suppose that $x, y > 0$, so that $x > \sqrt[3]{d}$, and then we have

$$(x - \sqrt[3]{d}y) = 1/(x^2 + xy\sqrt[3]{d} + y^2\sqrt[3]{d^2}) < 1/3y^2\sqrt[3]{d^2},$$

so that

$$|x/y - \sqrt[3]{d}| < 1/3y^3\sqrt[3]{d^2},$$

and this has only finitely many solutions by Roth's theorem (taking any $\epsilon < 1$).

(Note that Liouville's theorem just fails to prove this, as $\sqrt[3]{d}$ has degree 3.)

14. Sums of four squares

14.1. **The ring of quaternions.** We've used the arithmetic of the ring $\mathbb{Z}[i]$ to determine precisely which integers are the sum of two squares. On the other hand, it is a fact (first proved by Lagrange) that every positive integer

is the sum of *four* integer squares. This fact is connected with the arithmetic of a noncommutative ring, which we now describe.

Definition 14.1. The ring $\mathbb{H}$ of quaternions is the ring whose elements are formal sums $a + bi + cj + dk$, with $a, b, c, d \in \mathbb{R}$. Addition is given by the rule:

$$(a+bi+cj+dk)+(x+yi+zj+wk) = (a+x)+(b+y)i+(c+z)j+(d+w)k.$$

Multiplication is given by the rules:

$$i^2 = j^2 = k^2 = -1$$

$$ij = -ji = k$$

$$jk = -kj = i$$

$$ki = -ik = j$$

extended by $\mathbb{R}$-linearity and the distributive law.

Let $z = a + bi + cj + dk$ be a quaternion. The *conjugate* z^* of z is the quaternion $a - bi - cj - dk$. Note that if z and w are quaternions, then $(zw)^* = w^*z^*$.

The *norm* $N(z)$ of z is given by $N(z) = zz^* = a^2+b^2+c^2+d^2$. Note that this is a *real* number, and hence commutes with all elements of $\mathbb{H}$. Thus we have:

$$N(zw) = zw(zw)^* = zww^*z^* = zN(w)z^* = zz^*N(w) = N(z)N(w).$$

As was the case with $\mathbb{Z}[i]$, this gives an expression for the product of two sums of four squares as a sum of four squares. Explicitly, one has:

$$(a^2+b^2+c^2+d^2)(x^2+y^2+z^2+w^2) = N(a+bi+cj+dk)N(x+yi+zj+wk)$$

$$= N((a+bi+cj+dk)(x+yi+zj+wk))$$

$$= (ax-by-cz-dw)^2+(ay+bx+cw-dz)^2+(az-bw+cx+dy)^2+(aw+bz-cy+dx)^2.$$

In particular if m and n are integers that are representable as the sum of four integer squares, then so is their product mn. Thus to prove Lagrange's theorem it sufficies to prove that every prime is the sum of four integer squares.

14.2. **Proof of Lagrange's theorem.** Let p be a prime. If $p = 2$, or p is 1 mod 4, then we have already shown that p is a sum of two squares, hence also a sum of four squares. It thus remains to prove that primes congruent to 3 mod 4 are sums of four squares. We will do this via a descent argument, similar to that used in the proof of the two square theorem.

Lemma 14.2. *Let p be a prime congruent to 3 mod 4. Then there exist x and y such that $x^2 + y^2 + 1 \equiv 0$ (mod p).*

Proof. It suffices to find an integer a such that a is a square mod p and $a+1$ is not. Since -1 is not a quadratic residue mod p we would then have $-a-1$ a quadratic residue mod p. Taking x such that $x^2 \equiv a \bmod p$ and y such that $y^2 \equiv -a-1 \pmod p$ the claim would follow.

Suppose that we cannot do this. Then for each square a mod p, $a+1$ would also be a square mod p. In particular every congruence class mod p would be a square; since this doesn't happen we are done. □

Fix a prime congruent to 3 mod 4. By the lemma we can find x and y such that $x^2+y^2+1 = pr$ for some integer r. Since we only care about x and y mod p, we can further arrange that $|x|, |y| \leq \frac{p}{2}$. Then $r < p$. We are now ready to begin our descent:

Proposition 14.3. *Suppose that for some $p > r > 1$, we have x, y, z, w such that $x^2+y^2+z^2+w^2 = rp$. Then there exist x', y', z', w', r' integers with $1 \leq r' < r$ and $(x')^2+(y')^2+(z')^2+(w')^2 = r'p$.*

Proof. There are two cases we must treat separately. First suppose that r is even. Then either all of x, y, z, w have the same parity, or two of them are odd and two are even. Permuting x, y, z, w as necessary, we can assume x and y have the same parity, as do z and w. Then set:

$$\begin{aligned} x' &= \frac{x+y}{2} \\ y' &= \frac{x-y}{2} \\ z' &= \frac{z+w}{2} \\ w' &= \frac{z-w}{2} \end{aligned}$$

It is then easy to verify that $(x')^2+(y')^2+(z')^2+(w')^2 = \frac{r}{2}p$.

Now suppose that r is odd. Choose a, b, c, d such that $-\frac{r}{2} < a, b, c, d < \frac{r}{2}$ and $a \equiv x \pmod r$, $b \equiv y \pmod r$, etc. (we can get away with strict inequalities here because r is odd.)

Since $x^2+y^2+z^2+w^2 \equiv 0 \pmod r$, our congruences imply that $a^2+b^2+c^2+d^2 \equiv 0 \pmod r$. Write $a^2+b^2+c^2+d^2 = r'r$, and note that $r' < r$ since $a^2, b^2, c^2, d^2 < \frac{r^2}{4}$. On the other hand r' is nonzero (if it were zero, all of x, y, z, w would be divisible by r, and we would have $r^2 | rp$, hence $r|p$. This cannot happen since $1 < r < p$.) Thus we have $1 \leq r' < r$.

We then have:

$$r'r^2p = (a^2+b^2+c^2+d^2)(x^2+y^2+z^2+w^2)$$

$$= (ax+by+cz+dw)^2+(-ay+bx+cw-dz)^2+(-az-bw+cx+dy)^2+(-aw+bz-cy+dx)^2.$$

Note that $ax+by+cz+dw \equiv x^2+y^2+z^2+w^2 \equiv 0 \pmod r$. Similarly each of the other terms are congruent to zero mod r.

We thus have integers:

$$\begin{aligned} x' &= \frac{ax+by+cz+dw}{r} \\ y' &= \frac{-ay+bx+cw-dz}{r} \\ z' &= \frac{-az-bw+cx+dy}{r} \\ w' &= \frac{-aw+bz-vy+dx}{r} \end{aligned}$$

and $(x')^2+(y')^2+(z')^2+(w')^2 = r'p$ as desired. □

To complete the proof, one begins with $x^2+y^2+1 = pr$ and repeatedly applies the proposition until $r = 1$.

Remark 14.4. The descent in the proof of the two square theorem is really Euclid's algorithm for $\mathbb{Z}[i]$ in disguse. This descent is also Euclid's algorithm in a noncommutative setting. The associated ring is the ring of quaternions of the form $a+bi+cz+dw$, where either a, b, c, d are all integers, or a, b, c, d are all half-integers (that is, fractions of the form $\frac{r}{2}$ with r odd.)

14.3. **Sums of three squares.** At this point it's natural to ask which positive integers are the sum of *three* integer squares. This turns out to be much more difficult. In one direction, you showed on an earlier example sheet that no integer of the form $4^t(8k+7)$ (t, k integers) is a sum of three squares. Conversely, one has:

Theorem 14.5. *Every integer not of the form $4^t(8k+7)$ is a sum of three squares.*

The proof of this requires tools beyond the scope of the class, such as the Hasse principle for quadratic forms. One place to read about this is in Serre's *A course in arithmetic*.

15. Primes in arithmetic progressions

15.1. **Primes in arithmetic progressions.** A natural question to ask is how the primes are distributed mod n. It's easy to see that for any a with $(a, n) > 1$, there is at most one prime congruent to a mod n, so the best possible result is:

Theorem 15.1. *(Dirichlet) Given a positive integer n, and an integer a with $(a, n) = 1$, there are infinitely many primes congruent to a mod n.*

This was first proven by Dirichlet. The methods involved belong to analytic number theory; see for instance Serre's *A course in arithmetic* for a proof of this statement.

We will instead concern ourselves with special cases of this problem that can be approached by elementary methods.

15.2. **Elementary Results.** We first recall the proof that there are infinitely many primes; the structure of this proof will form a template for our arguments.

Theorem 15.2. *There are infinitely many primes.*

Proof. It suffices to construct, for any finite set S of primes, a prime not in S. Given such a set S, let $Q = 1 + \prod_{p \in S} p$. Then Q is divisible by at least one prime, but not by any prime in S. □

It's easy to adapt this proof to show that there are infinitely many primes congruent to 3 mod 4:

Theorem 15.3. *There are infinitely many primes congruent to* 3 *mod* 4.

Proof. Let S be a set of primes congruent to 3 mod 4, and let $Q = 2 + \prod_{p \in S} p^2$. Then Q is congruent to 3 mod 4, so Q is divisible by at least one prime congruent to 3 mod 4. On the other hand, Q is not divisible by any primes in S. □

A very similar argument works to show there are infinitely many primes congruent to 5 mod 6. Handling cases beyond that requires new ideas. For instance:

Lemma 15.4. *Let x be an even integer and let p be a prime dividing $x^2 + 1$. Then p is congruent to* 1 *mod* 4.

Proof. On the one hand p is clearly odd. On the other hand, $x^2 \equiv -1 \pmod{p}$, so -1 is a quadratic residue mod p. Hence p is 1 mod 4. □

Theorem 15.5. *There are infinitely many primes congruent to* 1 *mod* 4.

Proof. Let S be a finite set of primes congruent to 1 mod 4, and set $Q = 1 + \prod_{p \in S} p^2$. By the lemma, every prime dividing Q is congruent to 1 mod 4, but no prime in S divides Q. □

This suggests a strategy for proving there are infinitely many primes congruent to a mod n for some pair a, n: if we can find a polynomial P such that for any integer x, every prime dividing $P(nx)$ is congruent to a mod n, then we can try to mimic the proof that there are infinitely many primes congruent to 1 mod 4 to show that there are infinitely many primes congruent to a mod n. When $a = 1$ it turns out this is possible. In fact, it can be shown that this is possible if and only if $a^2 \equiv 1 \pmod{n}$, although this is a hard result (indeed, the "only if" direction involves proving a big generalisation of Dirichlet's theorem to number fields!). We will content ourselves with proving:

Theorem 15.6. *For any prime q, there are infinitely many primes congruent to* 1 *mod q.*

Definition 15.7. The qth cyclotomic polynomial Φ_q is the polynomial $(X^q - 1)/(X - 1) = X^{q-1} + \cdots + X + 1$.

Theorem 15.8. *Let $p \neq q$ be a prime, and let a be an integer. Then p divides $\Phi_q(a)$ if, and only if, a has exact order $q \pmod p$.*

Proof. a has exact order $q \pmod p$ if and only if $p|\Phi_q(a)$ but $p \nmid (a-1)$. But $\Phi_q(1) = q$ and $p \neq q$, so if $p|\Phi_q(a)$ then $p \nmid (a-1)$. □

Corollary 15.9. *Let $p \neq q$ be prime, and a an integer. If p divides $\Phi_q(a)$, then $p \equiv 1 \pmod q$.*

Proof. The order of $a \bmod p$ is q by the theorem above. Thus q divides $p-1$ by Fermat's little theorem. □

We are now in a position to prove:

Theorem 15.10. *Let q be a prime. There are infinitely many primes congruent to $1 \bmod q$.*

Proof. Let S be a (possibly empty) finite set of primes congruent to 1 mod q. Let R be the product of the primes in S, and consider $\Phi_q(qR) > 1$. By the corollary, every prime factor of $\Phi_q(qR)$ is either equal to q or is congruent to $1 \pmod q$. But $\Phi_q(qR) \equiv 1 \pmod{qR}$, so all of these prime factors ar $1 \pmod q$ and are not contained in S. □

Remark: in fact we could prove the theorem:

Theorem 15.11. *For any positive integer n, there are infinitely many primes congruent to $1 \bmod n$.*

in a similar way, by defining:

Definition 15.12. The nth cyclotomic polynomial Φ_n is the product:

$$\Phi_n = \prod_{1 \leq a < n; (a,n)=1} (x - e^{\frac{2\pi a i}{n}}).$$

In other words, Φ_n is the monic polynomial whose roots are the primitive nth roots of unity, all with multiplicity one.

The argument is then similar to the one we made in the prime case, but is a bit more fiddly.