Number Theory: A Contemporary Introduction

Pete L. Clark

Contents

CHAPTER 1

The Fundamental Theorem and Some Applications

1. Foundations

What is number theory?

This is a difficult question: number theory is an area, or collection of areas, of pure mathematics that has been studied for over two thousand years. As such, it means different things to different people. Nevertheless the question is not nearly as subjective as "What is truth?" or "What is beauty?": all of the things various people call number theory are related, in fact deeply and increasingly so over time.

If you think about it, it is hard to give a satisfactory definition of any area of mathematics that would make much sense to someone who has not taken one or more courses in it. One might say that analysis is the study of limiting processes, especially summation, differentiation and integration; that algebra is the study of algebraic structures like groups, rings and fields; and that topology is the study of topological spaces and continuous maps between them. But these descriptions function more by way of *dramatis personae* than actual explanations; less pretentiously, they indicate (some of) the *objects* studied in each of these fields, but they do not really tell us which properties of these objects are of most interest and which questions we are trying to answer about them. Such motivation is hard to provide in the abstract – much easier, and more fruitful, is to give examples of the types of problems that mathematicians in these areas are or were working on. For instance, in algebra one can point to the classification of finite simple groups, and in topology the Poincaré conjecture. Both of these are problems that had been open for long periods of time and have been solved relatively recently, so one may reasonably infer that these topics have been central to their respective subjects for some time.

What are the "objects" of number theory analogous to the above description? A good one sentence answer is that number theory is the study of the integers, i.e., the whole numbers and their negatives.

Of course this is not really satisfactory: astrology, accounting and computer science, for instance, could plausibly be described in the same way. What properties of the integers are we interested in?

The most succinct response seems to be that we are interested in the integers *as a ring*: namely, as endowed with the two fundamental operations of addition $+$ and multiplication $\cdot$ and – especially – the interactions between these two operations.

Let us elaborate. Consider first the non-negative integers – which, as is traditional, we will denote by $\mathbb{N}$ – endowed with the operation $+$. This is a very simple structure: we start with 0, the additive identity, and get every positive integer by repeatedly adding 1.[1] In some sense the natural numbers under addition are the simplest nontrivial algebraic structure.

Note that subtraction is not in general defined on the natural numbers: we would like to define $a - b = c$ in case $a = b + c$, but of course there is not always such a natural number c – consider e.g. $3 - 5$.

As you well know, there are two different responses to this: the first is to *formally extend* the natural numbers so that additive inverses always exist. In other words, for every positive integer n, we formally introduce a corresponding "number" $-n$ with the property that $n + (-n) = 0$. Although it is not *a priori* obvious that such a construction works – rather, the details and meaning of this construction were a point of confusion even among leading mathematicians for a few thousand years – nowadays we understand that it works to give a consistent structure: the integers $\mathbb{Z}$, endowed with an associative addition operation $+$, which has an identity 0 and for which each integer n has a unique additive inverse $-n$.

The second response is to record the relation between two natural numbers a and b such that $b - a$ exists as a natural number. Of course this relation is just that $a \leq b$. This is quite a simple relation on $\mathbb{N}$: indeed, for any pair of integers, we have either $a \leq b$ or $b \leq a$, and we have both exactly when $a = b$.[2]

Now for comparison consider the positive integers

$$\mathbb{Z}^+ = 1,\ 2, 3,\ \dots$$

under the operation of multiplication. This is a richer structure: whereas additively, there is a single building block: namely 1, the multiplicative building blocks are the prime numbers $2,\ 3,\ 5,\ 7, \dots$. Of course the primes are familiar objects, but the precise analogy with the additive case may not be as familiar, so let us spell it out carefully: just as subtraction is not in general defined on $\mathbb{N}$, division is not in general defined on $\mathbb{Z}^+$. On the one hand we can "formally complete" $\mathbb{Z}^+$ by adjoining multiplicative inverses, getting this time the positive rational numbers $\mathbb{Q}^+$. However, again one can view the fact that a/b is not always a positive integer as being intriguing rather than problematic, and we again consider the relation between two positive integers a and b that b/a be a positive integer: in other words, that there exist a positive integer c such that $b = a \times c$. In such a circumstance we say that a *divides* b, and write it as $a|b$.[3] The relation of divisibility is more complicated than the relation $\leq$ since divisibility is not a total ordering: e.g. $4 \not| 15$ and also $15 \not| 4$. What are we to make of this divisibility relation?

First, on a case-by-case basis, we *do* know how to determine whether $a \mid b$.

PROPOSITION 1.1. *(Division Theorem) For any positive integers n and d, there exist unique non-negative integers q and r with $0 \leq r < d$ and $n = qd + r$.*

[1] Here I am alluding to the fact that in the natural numbers, addition can be defined in terms of the "successor" operation $s(n) = n + 1$, as was done by the 19th century mathematical logician Giuseppe Peano. No worries if you have never heard of the Peano axioms – their importance lies in the realm of mathematical logic rather than arithmetic itself.

[2] That is to say, the relation $\leq$ on $\mathbb{N}$ is a linear, or total, ordering.

[3] Careful: $a|b \iff \frac{b}{a}$ is an integer.

This is a very useful tool, but it does not tell us the *structure* of $\mathbb{Z}^+$ under the divisibility relation. To address this, the primes inevitably come into play: there is a unique minimal element of $\mathbb{Z}^+$ under divisibility, namely 1 (in other words, 1 divides every positive integer and is the only positive integer with this property): it therefore plays the analogous role to 0 under $\leq$ on $\mathbb{N}$. In $\mathbb{N} \setminus 0$, the unique smallest element is 1. In $\mathbb{Z}^+ \setminus 1$ the smallest elements *with respect to the divisibility ordering* are the primes p. Given that the definition of a prime is precisely an integer greater than one divisible only by one and itself, this is clear. The analogue to repeatedly adding 1 is taking repeated powers of a single prime: e.g., 2, 2^2, 2^3, ... However, we certainly have more than one prime – in fact, as you probably know and we will recall soon enough, there are infinitely many primes – and this makes things more complicated. This suggests that maybe we should consider the divisibility relation one prime at a time.

So, for any prime p, let us define $a \mid_p b$ to mean that $\frac{b}{a}$ is a rational number which, when written in lowest terms, has denominator *not* divisible by p. For instance, $3 \mid_2 5$, since $\frac{5}{3}$, while not an integer, doesn't have a 2 in the denominator. For that matter, $3 \mid_p 5$ for all primes p different from 3, and this suggests the following:

PROPOSITION 1.2. *For any $a,\ b \in \mathbb{Z}^+$, $a|b \iff a \mid_p b$ for all primes p.*

PROOF. Certainly if $a|b$, then $a \mid_p b$ for all primes p. For the converse, write $\frac{b}{a}$ in lowest terms, say as $\frac{B}{A}$. Then $a \mid_p b$ iff A is not divisible by p. But the only positive integer which is not divisible by any primes is 1. □

In summary, we find that the multiplicative structure of $\mathbb{Z}^+$ is similar to the additive structure of $\mathbb{N}$, except that instead of there being one "generator" – namely 1 – such that every element can be obtained as some power of that generator, we have infinitely many generators – the primes – and every element can be obtained (uniquely, as we shall see!) by taking each prime a non-negative integer number of times (which must be zero for all but finitely many primes). This switch from one generator to infinitely many does not in itself cause much trouble: given

$$a = p_1^{a_1} \cdots p_n^{a_n} \cdots$$

and

$$b = p_1^{b_1} \cdots p_n^{b_n} \cdots$$

we find that $a \mid b$ iff $a \mid_p b$ for all p iff $a_i \leq b_i$ for all i. Similarly, it is no problem to multiply the two integers: we just have

$$ab = p_1^{a_1+b_1} \cdots p_n^{a_n+b_n} \cdots .$$

Thus we can treat positive integers under multiplication as vectors with infinitely many components, which are not fundamentally more complicated than vectors with a single component.

The "trouble" begins when we *mix* the additive and multiplicative structures. If we write integers in standard decimal notation, it is easy to add them, and if we write integers in the above "vector" factored form, it is easy to multiply them. But what is the prime factorization of $2^{137} + 3^{173}$? In practice, the problem of given an integer n, finding its prime power factorization (1) is extremely computationally difficult, to the extent that most present-day security rests on this difficulty.

It is remarkable how quickly we can find ourselves in very deep waters by asking apparently innocuous questions that mix additive and multiplicative structure. For instance, although in the multiplicative structure, each of the primes just rests "on its own axis" as a generator, in the additive structure we can ask where the primes occur with respect to the relation $\leq$. We do not have anything approaching a formula for p_n, and the task of describing the distribution of the p_n's inside $\mathbb{N}$ is a branch of number theory in and of itself (we will see a taste of it later on). For instance, consider the quantity $g(n) = p_{n+1} - p_n$, the "nth prime gap." For $n > 1$, the primes are all odd, so $g(n) \geq 2$. Computationally one finds lots of instances when $g(n)$ is exactly 2, e.g. $5, 7$, $11, 13$, and so forth: an instance of $g(n) = 2$ – equivalently, of a prime p such that $p + 2$ is also a prime – is called a *twin prime pair*. The trouble is that knowing the factorization of p tells us nothing[4] about the factorization of $p + 2$. Whether or not there are infinitely many twin primes is a big open problem in number theory.

(However it is not as open as when these notes were first written in 2007! At that time, for all that we knew it could have been the case that $\lim_{n\to\infty} g(n) = \infty$, i.e., for each constant C and all sufficiently large n, we have $g(n) = p_{n+1} - p_n > C$. This distressing possibility was disproved by Yitang Zhang in 2013: he showed that there are infinitely many n such that $g(n) < 70,000,000$. Later in 2013, James Maynard showed that $g(n) \leq 600$ for infinitely many n. Notice that if the 600 could be replaced with 2 we would have infinitely many twin primes. But we still don't know how to do that.)

It goes on like this: suppose we ask to represent numbers as a sum of two odd primes. Then such a number must be even and at least 6, and experimenting, one soon is led to guess that every even number at least 6 is a sum of two odd primes: this is known as Goldbach's Conjecture, and is about 400 years old. It remains unsolved.

(But again, there has been exciting recent progress. The Weak Goldbach Conjecture is that every odd integer $n \geq 9$ is the sum of three odd primes. It is easy to see that "Goldbach" implies "Weak Goldbach." In 1937 I.M. Vinogradov showed that every sufficiently large odd integer is the sum of three odd primes. In 2013 Harald Helfgott proved the Weak Goldbach Conjecture.)

There are many, many such easily stated unsolved problems which mix primes and addition: for instance, how many primes p are of the form $n^2 + 1$? Again, it is a standard conjecture that there are infinitely many, and it is wide open. Note that if we asked instead how many primes were of the form n^2, we would have no trouble answering – the innocent addition of 1 gives us terrible problems.

Lest you think we are just torturing ourselves by asking such questions, let me mention some amazing positive results:

THEOREM 1.3. *(Fermat, 12/25/1640) A prime $p > 2$ is of the form $x^2 + y^2$ iff it is of the form $4k + 1$.*

[4]Well, nothing except that $p + 2$ is not divisible by 2 for all $p > 2$.

This is, to my mind, the first beautiful theorem of number theory. It says that to check whether an odd prime satisfies the very complicated condition of being a sum of two (integer, of course!) squares, all we need to do is divide it by four: if its remainder is 1, then it is a sum of two squares; otherwise its remainder will be 3 and it will not be a sum of two squares.

THEOREM 1.4. *(Lagrange, 1770) Every positive integer is of the form* $x^2+y^2+z^2+w^2$.

THEOREM 1.5. *(Dirichlet, 1837) Suppose* a *and* b *are coprime positive integers (i.e., they are not both divisible by any integer* $n > 1$*). Then there are infinitely many primes of the form* $an + b$.

REMARK 1.6. Taking $a = 4,\ b = 1$, see that there are infinitely many primes of the form $4k + 1$, so in particular there are infinitely many primes which are a sum of two squares.

THEOREM 1.7. *(Green-Tao* [**GT08**]*) The primes contain arbitrarily long arithmetic progressions. That is: for any* $k \in \mathbb{Z}^+$ *there is a prime number* p *and a positive integer* d *such that*

$$p, p + d, p + 2d, \ldots, p + (k - 1)d$$

are all prime numbers.

We will see proofs of Theorems 1.3 and 1.4 in this course. To be more precise, we will give two different proofs of Theorem 1.3. The first theorem uses the observation that x^2+y^2 can be factored in the ring $\mathbb{Z}[i]$ of Gaussian integers as $(x+iy)(x-iy)$ and will be our jumping off point to the use of algebraic methods. There is an analogous proof of Theorem 1.4 using a noncommutative ring of "integral quaternions". This proof however has some technical complications which make it less appealing for in-class presentation, so we do not discuss it in these notes.[5] On the other hand, we will give parallel proofs of Theorems 1.3 and 1.4 using geometric methods. The proof of Theorem 1.5 is of a different degree of sophistication than any other proofs in this course. We do present a complete proof at the end of these notes, but one cannot pretend that this is undergraduate level material.

The proof of Theorem 1.7 is beyond the scope of this course.

Admission: In fact there is a branch of number theory which studies only the addition operation on subsets of $\mathbb{N}$: if A and B are two subsets of natural numbers, then by $A+B$ we mean the set of all numbers of the form $a+b$ for $a \in A$ and $b \in B$. For a positive integer h, by hA we mean the set of all h-fold sums $a_1 + \ldots + a_h$ of elements of A (repetitions allowed). There are plenty of interesting theorems concerning these operations, and this is a branch of mathematics called *additive number theory*. We will see a little bit of it towards the end of the course.

In fact Theorem 1.7 belongs to the subject of additive combinatorics. There is some irony that perhaps the single most celebrated result of that subject involves the multiplicative structure of $\mathbb{Z}$, but the practitioners of the field are well aware of this to say the least. We will say more about the issues here later when we talk about "density" and "substance" of integer sets.

[5]It was, in fact, the subject of a student project in the 2007 course.

2. The Fundamental Theorem (in $\mathbb{Z}$)

2.1. Existence of prime factorizations.

We had better pay our debts by giving a proof of the uniqueness of the prime power factorization. This is justly called the *Fundamental Theorem of Arithmetic*.

Let us first nail down the *existence* of a prime power factorization, although as mentioned above this is almost obvious:

PROPOSITION 1.8. *Every integer $n \geq 2$ is a product of primes $p_1 \cdots p_r$.*

PROOF. By induction on n. The base case, $n = 2$ is clear: $n = p_1$ is a prime. Suppose $n > 2$ and the result holds for all $2 \leq m < n$. Among all divisors $d > 1$ of n, the least is necessarily a prime, say p. So $n = pm$ and apply the result inductively to m. □

REMARK 1.9. If you are okay with the "empty product" – i.e., what we get by multiplying together 0 numbers – being 1, then the result extends also to $n = 1$. We will often find it superficially helpful to state things in this way, but there is certainly no content in it.

REMARK 1.10. Poposition 1.8 seemed obvious, and we proved it by induction. Formally speaking, just about any statement about the integers contain an appeal to induction at some point, since induction – or equivalently, the well-ordering principle that any nonempty subset of integers has a smallest element – is (along with a few much more straightforward axioms) their characteristic property. But induction proofs can be straightforward, tedious, or both. Often I will let you fill in such induction proofs; I will either just say "by induction" or, according to taste, present the argument in less formal noninductive terms. To be sure, sometimes an induction argument is nontrivial, and those will be given in detail.

In the above representation $n = p_1 \cdots p_r$ the same prime may of course occur more than once. Sometimes it is convenient to reorganize things: a **standard form** factorization of $n \geq 2$ is a factorization

$$n = p_1^{a_1} \cdots p_r^{a_r}$$

with $p_1 < \ldots < p_r$ primes and $a_1, \ldots, a_r$ positive integers. Any prime factorization yields a standard form prime factorization.

2.2. The fundamental theorem and Euclid's Lemma.

THEOREM 1.11. *The standard form factorization of a positive integer is* unique.

This is just a mildly laundered version of the more common statement: the factorization of a positive integer into primes is unique up to the order of the factors.

Theorem 1.11 was first stated and proved by Gauss in his *Disquisitiones Arithmeticae*. However, it is generally agreed that the result is "essentially" due to the ancient (circa 300 BC) Greek mathematician Euclid of Alexandria. Euclid proved:

THEOREM 1.12. *(Euclid's Lemma) Suppose p is prime, $a, b \in \mathbb{Z}^+$ and $p \mid ab$. Then $p \mid a$ or $p \mid b$.*

EXERCISE 1.1. Let p be a prime number, and let $a_1, \ldots, a_n \in \mathbb{Z}$. Show: if $p \mid a_1 \cdots a_n$, then $p \mid a_i$ for at least one $1 \leq i \leq n$.

Assuming the very easy Proposition 1.8, Theorems 1.11 and 1.12 are equivalent. From a strictly logical point of view two assertions are equivalent if they are both true or both false – or, if they range over a set of possible parameters then they are true for exactly the same values of those parameters. Since a theorem in mathematics is a true assertion, strictly speaking any two theorems are equivalent. But in common use the statement "Theorem A is equivalent to Theorem B" carries the connotation that it is much easier to deduce the truth of each theorem from the other than to prove either theorem. This is the case here.

Theorem 1.11 $\implies$ Theorem 1.12: Suppose for a contradiction that $p \mid ab$ but p does not divide either a or b. Writing out $a = \prod_i p_i^{a_i}$ and $b = \prod_j q_j^{b_j}$, our assumptions are equivalent to $p_i \neq p \neq q_j$ for all i, j. But then $ab = \prod p_i^{a_i} q_j^{a_j}$, and collecting this into standard form we get that no positive power of the prime p appears in the standard form factorization of ab. On the other hand, by assumption $p \mid ab$ so $ab = p \cdot m$, and then factoring m into primes we will get a standard form factorization of ab in which p does apear to some positive power, contradicting the uniqueness of the standard form prime factorization.

EXERCISE 1.2. Show that Theorem 1.11 implies the *Generalized Euclid's Lemma*: let $a, b, c \in \mathbb{Z}$. Suppose $a \mid bc$ and that no prime divides both a and b. Show: $a \mid c$.

Theorem 1.12 $\implies$ Theorem 1.11: Suppose $p_1 \cdots p_r = q_1 \cdots q_s$ are two prime factorizations of the same $n \geq 2$. Then $p_r \mid q_1 \cdots q_s$, so by Euclid's Lemma we have $p_r = q_j$ for some $1 \leq j \leq s$. After relabelling the q's if necessary we may assume that $p_r = q_s$ and cancel, getting

$$p_1 \cdots p_{r-1} = q_1 \cdots q_{s-1}.$$

Now we repeat the argument, cancelling p_{r-1} with (after relabelling if necessary) q_{s_1}. Eventually we get

$$1 = q_1 \cdots q_{s-r}.$$

But this means we have no more primes q, so $r = s$ and each p_i was equal to some q_j.

Therefore one way to prove Theorem 1.11 is to give Euclid's proof of Theorem 1.12. Euclid's proof goes by way of giving an explicit – and efficient – algorithm for finding the greatest common divisor of a pair of positive integers. This **Euclidean algorithm** can be put to a variety of uses in elementary number theory, so Euclid's proof is generally the one given in introductory courses. By making use of algebraic ideas it is possible to streamline Euclid's proof of Theorem 1.12 in a way which bypasses the algorithm: the idea is to show that the ring of integers has the property of being a **Principal Ideal Domain**, which is for a general ring a stronger result than the uniqueness of factorization into primes. In fact there is a third strategy, which directly proves Theorem 12.1. This proof, due to Lindemann [**Li33**] and Zermelo [**Z34**], is not sufficiently widely known. It is a nice instance of bypassing seemingly "necessary" machinery by sheer cleverness.

2.3. The Lindemann-Zermelo proof of the Fundamental Theorem.

We claim that the standard form factorization of a positive integer is unique. Assume not; then the set of positive integers which have at least two different standard form factorizations is nonempty, so has a least elment, say n, where:

$$n = p_1 \cdots p_r = q_1 \cdots q_s. \tag{1}$$

Here the p_i's and q_j's are prime numbers, not necessarily distinct from each other. However,we must have $p_1 \neq q_j$ for any j. Indeed, if we had such an equality, then after relabelling the q_j's we could assume $p_1 = q_1$ and then divide through by $p_1 = q_1$ to get a smaller positive integer $\frac{n}{p_1}$. By the assumed minimality of n, the prime factorization of $\frac{n}{p_1}$ must be unique: i.e., $r - 1 = s - 1$ and $p_i = q_i$ for all $2 \leq i \leq r$. But then multiplying back by $p_1 = q_1$ we see that we didn't have two different factorizations after all. (In fact this shows that for all i, j, $p_i \neq q_j$.)

In particular $p_1 \neq q_1$. Without loss of generality, assume $p_1 < q_1$. Then, if we subtract $p_1 q_2 \cdots q_s$ from both sides of (1), we get

$$m := n - p_1 q_2 \cdots q_s = p_1(p_2 \cdots p_r - q_2 \cdots q_s) = (q_1 - p_1)(q_2 \cdots q_s). \tag{2}$$

Evidently $0 < m < n$, so by minimality of n, the prime factorization of m must be unique. However, (2) gives two different factorizations of m, and we can use these to get a contradiction. Specifically, $m = p_1(p_2 \cdots p_r - q_2 \cdots q_s)$ shows that $p_1 \mid m$. Therefore, when we factor $m = (q_1 - p_1)(q_2 \cdots q_s)$ into primes, at least one of the prime factors must be p_1. But $q_2, \ldots, q_j$ are already primes which are different from p_1, so the only way we could get a p_1 factor is if $p_1 \mid (q_1 - p_1)$. But this implies $p_1 \mid q_1$, and since q_1 is also prime this implies $p_1 = q_1$. Contradiction!

T

2.4. Proof using ideals.

Now we turn things around by giving a direct proof of Euclid's Lemma. We (still!) do not follow Euclid's original proof, which employs the **Euclidean algorithm**, but rather a modernized version using ideals.

An **ideal** of $\mathbb{Z}$ is a nonempty subset I of $\mathbb{Z}$ such that $a,\ b \in I$ implies $a + b \in I$ and $a \in I, c \in \mathbb{Z}$ implies $ca \in I$.[6]

For any integer d, the set $(d) = \{nd \mid n \in \mathbb{Z}\}$ of all multiples of d is an ideal.

PROPOSITION 1.13. *Any nonzero ideal I of $\mathbb{Z}$ is of the form (d), where d is the least positive element of I.*

PROOF. Suppose not: then there exists an element n which is not a multiple of d. Applying the Division Theorem (Proposition B.11), we may write $n = qd + r$ with $0 < r < d$. Since $d \in I$, $qd \in I$ and hence $r = n - qd \in I$. But r is positive and smaller than d, a contradiction. □

Existence of gcd's: Let a and b be two nonzero integers. An integer d is said to be a **greatest common divisor** of a and b if

(GCD1) $d \mid a$ and $d \mid b$.
(GCD2) If $e \mid a$ and $e \mid b$ then $e \mid d$.

[6]We hope that the reader recognizes this as a special case of an ideal in a commutative ring.

Note well that this is (at least apparently) different from the definition of greatest common divisor one learns in school: in the set of all common divisors of a and b, d is defined to be a divisor which is divisible by every other divisor, not a divisor which is numerically largest. In particular, unlike the school definition, it is not obvious that greatest common divisors exist! However:

PROPOSITION 1.14. *For $a, b \in \mathbb{Z}$, not both zero, the set $I_{a,b} = \{xa + yb \mid x, y \in \mathbb{Z}\}$ is a nonzero ideal. Its positive generator d has the following property:*

$$e|a \ \& \ e|b \iff e|d, \tag{3}$$

and is therefore a greatest common divisor of a and b.

PROOF. It is easy to see that the set $I_{a,b}$ is closed under addition and under multiplication by all integers, so it is an ideal. By the previous result, it is generated by its smallest positive element, say $d = Xa + Yb$.

Now, suppose $e|d$. Then, since $a \in (d)$, $(a) \subset (d)$ and thus $d|a$ (to contain is to divide) and by transitivity $e|a$; similarly $e|b$. (In fact we made a bigger production of this than was necessary: we could have said that d is a multiple of e, and a and b are multiples of d, so of course a and b are multiples of e. This is the easy direction.) Conversely, suppose that $e|a$ and $e|b$ (so e is a common divisor of a and b). Then $e \mid Xa + Yb = d$. (Since d could be smaller than a and b – e.g. $a = 17$, $b = 10^{10}$, $d = 1$, this is the nontrivial implication.) □

COROLLARY 1.15. *If a and b are integers, not both zero, then for any integer m there exist integers x and y such that*

$$xa + yb = m \gcd(a, b).$$

PROOF. This follows immediately from the equality of ideals $I_{a,b} = (\gcd(a, b))$: the left hand side is an arbitrary element of $I_{a,b}$ and the right hand side is an arbitrary element of $(\gcd(a, b))$. □

An important special case is when $\gcd(a, b) = 1$ – we say a and b are **relatively prime**. The corollary then asserts that for any integer m, we can find integers x and y such that $xa + yb = m$.

Indeed we can use this to prove Euclid's Lemma (Theorem **??**): if $p \mid ab$ and p does not divide a, then the greatest common divisor of p and a must be 1. Thus there are integers x and y such that $xa + yp = 1$. Multiplying through by b we get $xab + ypb = b$. Since $p \mid xab$ and $p \mid ypb$, we conclude $p \mid b$. This completes the proof of the Fundamental Theorem of Arithmetic.

3. Some examples of failure of unique factorization

The train of thought involved in proving the fundamental theorem is quite subtle. The first time one sees it, it is hard to believe that such complications are necessary: is it not "obvious" that the factorization of integers into primes is unique?

It is not obvious, but rather familiar and true. The best way to perceive the non-obviousness is to consider new and different contexts.

Example: let $\mathbb{E}$ denote the set of even integers.[7] Because this is otherwise known as the ideal $(2) = 2\mathbb{Z}$, it has a lot of structure: it forms a group under addition, and there is a well-defined multiplication operation satisfying all the properties of a ring except one: namely, there is no 1, or multiplicative identity. (A ring without identity is sometimes wryly called a *rng*, so the title of this section is not a typo.)

Let us consider factorization in $\mathbb{E}$: in general, an element x of some structure should be prime if every factorization $x = yz$ is "trivial" in some sense. However, in $\mathbb{E}$, since there is no 1, there are no trivial factorizations, and we can define an element x of $\mathbb{E}$ to be prime if it cannot be written as the product of two other elements of $\mathbb{E}$. Of course this is a new notion of prime: 2 is a conventional prime and also a prime of $\mathbb{E}$, but clearly none of the other conventional primes are $\mathbb{E}$-prime. Moreover there are $\mathbb{E}$-primes which are not prime in the usual sense: e.g., 6 is $\mathbb{E}$-prime. Indeed, it is not hard to see that an element of $\mathbb{E}$ is an $\mathbb{E}$-prime iff it is divisible by 2 but not by 4.

Now consider

$$36 = 2 \cdot 18 = 6 \cdot 6.$$

Since 2, 18 and 6 are all divisible by 2 and not 4, they are $\mathbb{E}$-primes, so 36 has two different factorizations into $\mathbb{E}$-primes.

This example begins to arouse our skepticism about unique factorization: it is not, for instance, inherent in the nature of factorization that factorization into primes must be unique. On the other hand, the rng $\mathbb{E}$ is quite artificial: it is an inconveniently small substructure of a better behaved ring $\mathbb{Z}$. Later we will see more distressing examples.

Example 2: Let $R_\circ = \mathbb{R}[\cos\theta, \sin\theta]$ be the ring of real trigonometric polynomials: i.e., the ring whose elements are polynomial expressions in $\sin\theta$ and $\cos\theta$ with real coefficients. We view the elements as functions from $\mathbb{R}$ to $\mathbb{R}$ and add and multiply them pointwise.

Of course this ring is not isomorphic to the polynomial ring $\mathbb{R}[x, y]$, since we have the Pythagorean identity $\cos^2\theta + \sin^2\theta = 1$. It is certainly plausible – and can be shown to be true – that all polynomial relations between the sine and cosine are consequences of this one relation, in the sense that $R_\circ$ is isomorphic to the quotient ring $\mathbb{R}[x, y]/(x^2 + y^2 - 1)$.

Now consider the basic trigonometric identity

$$(\cos\theta)(\cos\theta) = (1 + \sin\theta)(1 - \sin\theta). \tag{4}$$

It turns out that $\cos\theta$, $1 + \sin\theta$ and $1 - \sin\theta$ are all irreducible elements in the ring $R_\circ$.[8] Moreover, if $f, g \in R_\circ$ are associates – i.e., there is an invertible element $u \in R$ such that $g = uf$, then u does not vanish at any point on the unit cricle, and thus the subsets of the unit circle on which f and g vanish are the same. But the subsets of the unit circle on which $\cos\theta$, $1 + \sin\theta$ and $1 - \sin\theta$ vanish are $\{\pm\frac{\pi}{2}\}$, $\{\pi\}$ and $\{0\}$, repsectively, so so all three of these elements are nonassociate, and therefore (34) exhibits two different factorizations into irreducible elements! Thus,

[7]This example is taken from Silverman's book. In turn Silverman took it, I think, from Harold Stark's introductory number theory text. Maybe it is actually due to Stark...

[8]To be sure, this claim requires proof, which is being omitted here!

in a sense, the failure of unique factorization in $R_\circ$ is the explanation for the subject of trigonometric identities!

To see how subtle the issue of unique factorization can be, consider now the ring

$$C_\circ = \mathbb{C}[\cos\theta, \sin\theta]$$

of trigonometric polynomials with complex coefficients. But the classic "Euler identity"

$$e^{i\theta} = \cos\theta + i\sin\theta$$

shows that $e^{i\theta}$ is an element of $C_\circ$, and conversely, both the sine and cosine functions are expressible in terms of $e^{i\theta}$:

$$\cos\theta = \frac{1}{2}\left(e^{i\theta} + \frac{1}{e^{i\theta}}\right),$$

$$\sin\theta = \frac{1}{2i}\left(e^{i\theta} - \frac{1}{e^{i\theta}}\right).$$

Thus $C_\circ = \mathbb{C}[e^{i\theta}, \frac{1}{e^{i\theta}}]$. Now the ring $\mathbb{C}[e^{it}]$ is isomorphic to the polynomial ring $\mathbb{C}[T]$, so $C_\circ$ is, up to isomorphism, obtained from $\mathbb{C}[T]$ by adjoining T^{-1}. Recall that $\mathbb{C}[t]$ is a principal ideal domain (PID). Finally, if R is any PID with fraction field K, and S is any ring such that $R \subset S \subset K$ – i.e., any ring obtained by adjoining to R the multiplicative inverses of each of some set of nonzero elements of R – then it can be shown that S is also a PID, hence in particular a unique factorization domain.

The foregoing discussion has been quite brief, with no pretense of presenting a complete argument. For a more detailed discussion, I highly recommend [**Tr88**]. A more sophisticated presentation can be found in [**Cl09**, Thm. 12].

4. Consequences of the fundamental theorem

The second proof of the fundamental theorem develops material which is very useful in its own right. Let us look at some of it in more detail:

4.1. Applications of the prime power factorization.

There are certain functions of n which are most easily defined in terms of the prime power factorization. This includes many so-called **arithmetic functions** that we will discuss a bit later in the course. But here let us give some basic examples. First, let us write the prime power factorization as

$$n = \prod_i p_i^{a_i},$$

where p_i denotes the ith prime in sequence, and a_i is a non-negative integer. This looks like an infinite product, but we impose the condition that $a_i = 0$ for all but finitely many i,[9] so that past a certain point we are just multiplying by 1. The convenience of this is that we do not need different notation for the primes dividing some other integer.

[9]In fact, this representation is precisely analogous to the expression of $(\mathbb{Z}^\cdot, \cdot) = (\mathbb{N}, +)^\infty$ of problem G1).

Now suppose we have two such factored positive integers

$$a = \prod_i p_i^{a_i},$$
$$b = \prod_i p_i^{b_i}.$$

Then we can give a simple and useful formula for the gcd and the lcm. Namely, the greatest common divisor of a and b is

$$\gcd(a,b) = \prod_i p_i^{\min(a_i,b_i)},$$

where $\min(c,d)$ just gives the smaller of the two integers c and d (and, of course, the common value $c = d$ when they are equal). More generally, we have that, writing out two integers a and b in factored form above, we have that $a \mid b \iff a_i \leq b_i$ for all i. In fact this is exactly the statement that $a|b \iff a|_p b$ for all p that we expressed earlier.

We often (e.g. now) find ourselves wanting to make reference to the a_i in the prime power factorization of an integer a. The a_i is the highest power of p_i that divides a. One often says that $p_i^{a_i}$ *exactly divides* a, meaning that $p_i^{a_i}|a$ and $p_i^{a_i+1}$ does not. So let us define, for any prime p, $\mathrm{ord}_p(a)$ to be the highest power of p that divides a: equivalently:

$$n = \prod_i p_i^{\mathrm{ord}_{p_i}(n)}.$$

Notice that ord_p is reminiscent of a logarithm to the base p: in fact, that's exactly what it is when $n = p^a$ is a power of p only: $\mathrm{ord}_p(p^a) = a$. However, for integers n divisible by some prime $q \neq p$, $\log_p(n)$ is nothing nice – in fact, it is an irrational number – whereas $\mathrm{ord}_p(n)$ is by definition always a non-negative integer. In some sense, the beauty of the functions ord_p is that they allow us to "localize" our attention at one prime at a time: every integer n can be written as $p^r \cdot m$ with $\gcd(m,p) = 1$, and the ord_p just politely ignores the m: $\mathrm{ord}_p(p^r \cdot m) = \mathrm{ord}_p(p^r) = r$.

This is really just notation, but it is quite useful: for instance, we can easily see that for all p,

$$\mathrm{ord}_p(\gcd(a,b)) = \min(\mathrm{ord}_p(a), \mathrm{ord}_p(b));$$

this just says that the power of p which divides the gcd of a and b should be the largest power of p which divides both a and b. And then a positive integer n is determined by all of its $\mathrm{ord}_p(n)$'s via the above equation.

Similarly, define the least common multiple $\mathrm{lcm}(a,b)$ of positive integers a and b to be a positive integer m with the property that $a|e$ & $b|e \implies m|e$. Then essentially the same reasoning gives us that

$$\mathrm{ord}_p(\mathrm{lcm}(a,b)) = \max(\mathrm{ord}_p(a), \mathrm{ord}_p(b)),$$

and then that

$$\mathrm{lcm}(a,b) = \prod_p p^{\max(\mathrm{ord}_p(a), \mathrm{ord}_p(b))}.$$

We can equally well define ord_p on a negative integer n: it is again the largest power i of p such that $p_i | n$. Since multiplying by -1 doesn't change divisibility in any way, we have that $\mathrm{ord}_p(n) = \mathrm{ord}_p(-n)$. Note however that $\mathrm{ord}_p(0)$ is slightly problematic – every p^i divides 0: $0 \cdot p^i = 0$ – so if we are going to define this at all it would make sense to put $\mathrm{ord}_p(0) = \infty$.

We do lose something by extending the ord functions to negative integers: namely, since for all p, $\mathrm{ord}_p(n) = \mathrm{ord}_p(-n)$, the ord functions do not allow us to distinguish between n and $-n$. From a more abstract algebraic perspective, this is because n and $-n$ generate the same ideal (are **associates**; more on this later), and we make peace with the fact that different generators of the same ideal are more or less equivalent when it comes to divisibility. However, in $\mathbb{Z}$ we do have a remedy: we could define a map $\mathrm{ord}_{-1} : \mathbb{Z} \setminus \{0\} \to \pm 1$ such that $\mathrm{ord}_{-1}(n) = +1$ if $n > 0$ and -1 if $n < 0$. Then -1 acts as a "prime of order 2," in contrast to the other "infinite order primes," and we get a corresponding unique factorization statement.[10] But although there is some sense to this, we will not adopt it formally here.

PROPOSITION 1.16. *For p a prime and m and n integers, we have:*
a) $\mathrm{ord}_p(mn) = \mathrm{ord}_p(m) + \mathrm{ord}_p(n)$.
b) $\mathrm{ord}_p(m+n) \geq \min(\mathrm{ord}_p(m), \mathrm{ord}_p(n))$.
c) If $\mathrm{ord}_p(m) \neq \mathrm{ord}_p(n)$, $\mathrm{ord}_p(m+n) = \min(\mathrm{ord}_p(m), \mathrm{ord}_p(n))$.

We leave these as exercises: suitably decoded, they are familiar facts about divisibility. Note that part a) says that ord_p is some sort of *homomorphism* from $\mathbb{Z} \setminus \{0\}$ to $\mathbb{Z}$. However, $\mathbb{Z} \setminus \{0\}$ under multiplication is not our favorite kind of algebraic structure: it lacks inverses, so is a monoid rather than a group. This perhaps suggests that we should try to extend it to a map on the nonzero rational numbers $\mathbb{Q}^\times$ (which, if you did problem G1), you will recognize as the group completion of $\mathbb{Z} \setminus \{0\}$; if not, no matter), and this is no sooner said than done:

For a nonzero rational number $\frac{a}{b}$, we define

$$\mathrm{ord}_p(\frac{a}{b}) = \mathrm{ord}_p(a) - \mathrm{ord}_p(b).$$

In other words, powers of p dividing the numerator count positively; powers of p dividing the denominator count negatively. There is something to check here, namely that the definition does not depend upon the choice of representative of $\frac{a}{b}$. But it clearly doesn't:

$$\mathrm{ord}_p(\frac{ac}{bc}) = \mathrm{ord}_p(ac) - \mathrm{ord}_p(bc)$$

$$= \mathrm{ord}_p(a) + \mathrm{ord}_p(c) - \mathrm{ord}_p(b) - \mathrm{ord}_p(c) = \mathrm{ord}_p(a) - \mathrm{ord}_p(b) = \mathrm{ord}_p(\frac{a}{b}).$$

So we get a map

$$\mathrm{ord}_p : \mathbb{Q}^\times \to \mathbb{Z}$$

which has all sorts of uses: among other things, we can use it to recognize whether a rational number x is an integer: it will be iff $\mathrm{ord}_p(x) \geq 0$ for all primes p.

Example: Let us look at the partial sums S_i of the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$. The first partial sum $S_1 = 1$ – that's a whole number. The second one is $S_2 = 1 + \frac{1}{2} = \frac{3}{2}$

[10] This perspective is due to J.H. Conway.

which is not. Then $S_3 = 1 + \frac{1}{2} + \frac{1}{3} = \frac{11}{6}$ is not an integer either; neither is $S_4 = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12}$.

It is natural to ask whether *any* partial sum S_n for $n \geq 1$ is an integer. Indeed, this is a standard question in honors math classes because...well, frankly, because it's rather hard.[11] But using properties of the ord function we can give a simple proof. The first step is to look carefully at the data and see if we can find a pattern. (This is, of course, something to do whenever you are presented with a problem whose solution you do not immediately know. Modern presentations of mathematics – including, alas, these notes, to a large extent – often hide this experimentation and discovery process.) What we see in the small partial sums is that not only are they not integers, they are all not integers for "the same reason": there is always a power of 2 in the denominator.

So what we'd like to show is that for all $n \geq 1$, $\mathrm{ord}_2(S_n) < 0$. It is true for $n = 2$; moreover we don't have to do the calculation for $n = 3$: since $\mathrm{ord}_2(\frac{1}{3}) = 0 \neq \mathrm{ord}_2(S_2)$, we must have $\mathrm{ord}_2(S_2 + \frac{1}{3}) = \min(\mathrm{ord}_2(S_2), \mathrm{ord}_2(S_3)) = -1$. And then we get $\frac{1}{4}$, which 2-order -2, which is different from $\mathrm{ord}_2(S_3)$, so again, using that when we add two rational numbers with different 2-orders, the 2-order of the sum is the smaller of the 2 2-orders, we get that $\mathrm{ord}_2(S_4) = -2$. Excitedly testing a few more values, we see that this pattern continues: $\mathrm{ord}_2(S_n)$ and $\mathrm{ord}_2(\frac{1}{n+1})$ are always different; if only we can show that this always holds, this will prove the result. In fact one can say even more: one can *precisely* what $\mathrm{ord}_2(S_n)$ is as a function of n and thus see in particular that it is always negative. I will leave the final observation and proof to you – why should I steal your fun?

4.2. Linear Diophantine equations.

Recall that one of the two main things we agreed that number theory is about was solving Diophantine equations, i.e., looking for solutions over $\mathbb{Z}$ and/or over $\mathbb{Q}$ to polynomial equations. Certainly we saw some primes in the previous section; now we solve the simplest class of Diophantine equations, namely the linear ones.

Historical remark: as I said before, nowadays when someone says Diophantine equation, they mean that we are interested either in solutions over $\mathbb{Z}$ or solutions over $\mathbb{Q}$, or both. Diophantus himself considered positive rational solutions. Nowadays the restriction to positive numbers seems quite artificial (and I must wonder whether Diophantus massaged his equations so as to get positive rather than negative solutions); it also makes things quite a bit more difficult: it stands to reason that since equations become easier to solve if we allow ourselves to divide numbers, correspondingly they become more difficult if we do not allow subtraction!

This also means that the term "Linear Diophantine equation" is, strictly speaking, an anachronism. If you want to solve any number of linear equations with coefficients in $\mathbb{Q}$, then – since $\mathbb{Q}$ is a field – you are just doing linear algebra, which works equally well over $\mathbb{Q}$ as it does over $\mathbb{R}$ or $\mathbb{C}$. For instance, suppose we want to

[11] When I first got assigned this problem (my very first semester at college), I found – or looked up? – some quite elaborate solution which used, in particular, **Bertrand's Postulate** that for $n > 1$ there is always a prime p with $n < p < 2n$. (This was proven in the latter half of the 19th century by Cebyshev. One of Paul Erdős' early mathematical triumphs was an elegant new proof of this result.)

solve the equation

$$ax + by = c$$

in rational numbers, where a and b are nonzero rational numbers and c is any rational number. Well, it's not much fun, is it? Let x be any rational number at all, and solve for y:

$$y = \frac{c - ax}{b}.$$

Speaking more geometrically, any line $y = mx + b$ in the plane passing through one rational point and with rational slope – roughly speaking, with m and b rational – will have lots of rational solutions: one for every rational choice of x.

So for Diophantus, the first interesting example was quadratic polynomial equations. Indeed, after this section, the quadratic case will occupy our interest for perhaps the majority of the course.

However, over $\mathbb{Z}$ things are never so easy: for instance, the equation

$$3x + 3y = 1$$

clearly does not have an integer solution, since no matter what integers x and y we choose, $3x + 3y$ will be divisible by y. More generally, if a and b have a common divisor $d > 1$, then it is hopeless to try to solve

$$ax + by = 1.$$

But this is the only restriction, and indeed we saw this before: en route to proving the fundamental theorem, we showed that for any integers a and b, not both zero, then $\gcd(a, b)$ generates the ideal $\{xa + yb \mid x,\ y \in \mathbb{Z}\}$, meaning that for any integer m, the equation

$$ax + by = m \gcd(a, b)$$

has solutions in x and y. In other words, we can solve

$$ax + by = n$$

if n is a multiple of the gcd of a and b. By the above, it is also true that we can only solve the equation if n is a multiple of the gcd of x and y – the succinct statement is the *equality* of ideals $I_{a,b} = (\gcd(a, b))$ – so we have (and already had, really) the following important result.

THEOREM 1.17. *For fixed $a, b \in \mathbb{Z}$, not both zero, and any $m \in \mathbb{Z}$, the equation*

$$ax + by = m$$

has a solution in integers (x, y) iff $\gcd(a, b) \mid m$.

In particular, if $\gcd(a, b) = 1$, then we can solve the equation for any integer m. The fundamental case is to solve

$$ax + by = 1,$$

because if we can find such x and y, then just by multiplying through by m we can solve the general equation.

This is a nice result, but it raises two further questions. First, we found one

solution. Now what can we say about *all* solutions?[12] Second, given that we know that solutions exist, how do we actually find them?

EXAMPLE 1.18. *We are claiming that $3x + 7y = 1$ has an integer solution. What could it be? Well, a little experimentation yields $x = -2$, $y = 1$. Is this the only solution? Indeed not: we could add 7 to x and the sum would increase by 21, and then subtract 3 from y and the sum would decrease by 21. This leads us to write down the family of solutions $x_n = -2 + 7n$, $y_n = 1 - 3n$. Are there any more? Well, we have found one integral solutions whose x-coordinates are evenly spaced, 7 units apart from each other. If there is any other solution $3X + 7Y = 1$, there must be some n such that $0 < X - x_n < 7$. This would give a solution $3(X - x_n) = -7(Y - y_n)$ with $0 < X - x_n < 7$. But this is absurd: the left hand side would therefore be prime to 7, whereas the right hand side is divisible by 7. So we evidently found the general solution.*

The above argument does not, of course, use any special properties of 3 and 7: with purely notational changes it carries over to a proof of the following result.

THEOREM 1.19. *For a and b coprime positive integers, the general integral solution to $xa + yb = 1$ is $x_n = x_0 + nb$, $y_n = y_0 - na$, where $x_0a + y_0b = 1$ is any particular solution guaranteed to exist by Theorem 1.17.*

However, let us take the opportunity to give a slightly different reformulation and reproof of Theorem 1.19. We will work in slightly more generality: for fixed, relatively prime nonzero integers a and b and a variable integer N, consider all integral solutions of the equation

$$ax + by = N \tag{5}$$

To borrow terminology from other areas of mathematics,[13] (5) is **linear and inhomogeneous** in x and y. What this means is that the left hand side is an expression which is linear in x and y but the right-hand side is nonzero. There is an associated **homogeneous linear equation**:

$$ax + by = 0 \tag{6}$$

Here we are saying something quite basic in a fancy way: the real solutions of (6) form a line through the origin in $\mathbb{R}^2$, with slope $m = \frac{-a}{b}$. But the set of integer solutions to (6) also has a nice algebraic structure: if (x_1, y_1), (x_2, y_2) are any two integer solutions and C is any integer, then since

$$a(x_1 + x_2) + b(y_1 + y_2) = (ax_1 + by_1) + (ax_2 + by_2) = 0 + 0 = 0,$$
$$a(Cx_1) + b(Cy_2) = C(ax_1 + by_1) = C \cdot 0 = 0,$$

both the sum $(x_1, y_1) + (x_2, y_2)$ and the integer multiple $C(x_1, y_1)$ are solutions. To be algebraically precise about it, the set of integer solutions to (6) forms a subgroup of the additive group of the one-dimensional $\mathbb{R}$-vector space of all real solutions.

Now we claim that it is easy to solve the homogeneous equation directly. The $\mathbb{Q}$-solutions are clearly $\{(x, \frac{-a}{b}x) \mid x \in \mathbb{Q}\}$. And, since a and b are relatively prime, in order for x and $\frac{-a}{b}x$ to both be integers, it is necessary and sufficient that x

[12] Diophantus was for the most part content with finding a single solution. The more penetrating inquiry into the set of all solutions was apparently first made by Fermat.

[13] Especially, from the elementary theory of differential equations.

itself be an integer and that it moreover be divisible by b. Therefore the general integral solution to the homogeneous equation is $\{(nb, -na) \mid n \in \mathbb{Z}\}$.

Now we make the fundamental observation about solving inhomogeneous linear equations in terms of the associated homogeneous linear equation. We claim that if (x_0, y_0) is any one solution to the inhomogeneous equation (5) and $(x_n, y_n) = (nb, -na)$ is the general solution to the associated homogeneous equation (6), then the general solution to the inhomogeneous equation is $(x_0, y_0) + (x_n, y_n)$. Let's check this. On the one hand, we have

$$a(x_0 + x_n) + b(y_0 + y_n) = (ax_0 + by_0) + (ax_n + by_n) = N + 0 = N,$$

so these are indeed solutions to the inhomogeneous equation. On the other hand, if (x, y) and (x', y') are any two solutions to the inhomogeneous equation, then, by a very similar computation, their difference $(x - x', y - y')$ is a solution to the homogeneous equation.

In other words the set of all solutions to the inhomogeneous equation is simply a **translate** of the abelian group of all solutions to the homogeneous equation. Thus, since the solutions to the homogeneous equation are simply a set of points along the line with distance $\sqrt{a^2 + b^2}$ between consecutive solutions, the same holds for **all** the inhomogeneous equations, independent of N.

Remark aside: At the cost of introducing some further fancy terminology, the discussion can be summarized by saying that the solution set to the inhomogeneous equation is a **principal homogeneous space** for the commutative group of solutions to the homogeneous equation. The general meaning of this is in terms of group actions on sets: let G be a group, X a set, and $\bullet : G \times X \to X$ an action of G on X. (We are assuming familiarity with this algebraic concept only to make the present digression. It will not be needed in the rest of the course.) Then we say that X is a principal homogeneous space for G if the action is simply transitivie: for all $x, y \in X$, there exists a unique element g of G such that $g \cdot x = y$.

To look back this homogeneous/inohomogeneous argument, what it *doesn't* give us is any particular solution to the inhomogeneous equation. (Indeed, so far as this abtract reasoning goes, such a solution might not exist: according to the definition we gave for principal homogeneous space, taking $X = \varnothing$ gives a principal homogeneous space under any group G!) To get this in any given case we can use Euclid's algorithm, but in thinking about things in general it is useful to acknowledge a certain amount of fuzziness in the picture: we can only say where any particular solution will be located on the line to within an accuracy of $d = \sqrt{a^2 + b^2}$.

What is interesting is that we can use these seemingly very primitive geometric ideas to extract useful information about a more difficult problem. Namely, let us now suppose that a, b, N are all positive integers, and we seek to solve the linear Diophantine equation

$$ax + by = N$$

in positive integers (x, y). Then the geometric picture shows right away that we are interested in the intersection of the infinite family of all integral solutions with

the first quadrant of $\mathbb{R}^2$. More precisely, we have a line segment L_N which joins $(0, \frac{N}{b})$ to $(\frac{N}{a}, 0)$, and we are asking whether there are integer solutions on L_N.

Notice that the length of L_N is

$$\ell_N = \sqrt{\left(\frac{N}{a}\right)^2 + \left(\frac{N}{b}\right)^2} = N\sqrt{\frac{1}{a^2} + \frac{1}{b^2}} = N\frac{\sqrt{a^2+b^2}}{ab} = \left(\frac{d}{ab}\right)N.$$

Thus when N is small, L_N is a very small line segment, and since consecutive integral solutions on the line are spaced d units apart, it is by no means guaranteed that there are any integral solutions on L_N. For instance, since $ax+by \geq a+b \geq 2$, there is no positive integral solution to $ax + by = 1$. But since L_N grows linearly with N and d is independent of N, when N is sufficiently large we must have some integral points on L_N. In fact this must happen as soon as $\ell_N > d$.[14] By similar reasoning, the number of solutions must be extremely close to $\frac{\ell_n}{d} = \frac{N}{ab}$. Precisely:

THEOREM 1.20. *Let $a, b \in \mathbb{Z}^+$ be relatively prime, and let $N \in \mathbb{Z}^+$.*
a) If $N > ab$, then there exist positive integers x, y such that $ax + by = N$.
b) Let $\mathcal{N}_N$ be the number of positive integral solutions (x, y) to $ax + by = N$. Then

$$\lfloor \frac{N}{ab} \rfloor - 1 \leq \mathcal{N}_N \leq \lfloor \frac{N}{ab} \rfloor + 1.$$

We leave the details of the proof of Theorem 1.20 to the interested reader.

It turns out that the lower bound on N in part a) is of the right order of magnitude, but is never sharp: for instance, if $a = 2$, $b = 3$, then the theorem asserts $2x + 3y = N$ has a positive integral solution if $N > 6$, whereas pure thought shows that it suffices to take $N \geq 2$. The sharp lower bound is known (in terms of a and b, of course) and is a result of J.J. Sylvester [**Sy84**].

5. Some Irrational Numbers

PROPOSITION 1.21. *The square root of 2 is irrational.*

PROOF. Suppose not: then there exist integers a and $b \neq 0$ such that $\sqrt{2} = \frac{a}{b}$, meaning that $2 = \frac{a^2}{b^2}$. We may assume that a and b have no common divisor – if they do, divide it out – and in particular that a and b are not both even.

Now clear denominators:

$$a^2 = 2b^2.$$

So $2 \mid a^2$. It follows that $2 \mid a$. Notice that this is a direct consequence of Euclid's Lemma – if $p \mid a^2$, $p \mid a$ or $p \mid a$. On the other hand, we can simply prove the contrapositive: if a is odd, then a^2 is odd. By the Division Theorem, a number is odd iff we can represent it as $a = 2k + 1$, and then we just check: $(2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$ is indeed again odd. So $a = 2A$, say. Plugging this into the equation we get

$$(2A)^2 = 4A^2 = 2b^2,\ b^2 = 2A^2,$$

so $2 \mid b^2$ and, as above, $2 \mid b$. Thus 2 divides both a and b: contradiction. □

[14] To understand the reasoning here, imagine that you know that a certain bus comes once every hour at a fixed time – i.e., at a certain number of minutes past each hour – but you don't know exactly what that fixed time is. Nevertheless, if you wait for any full hour, you will be able to catch the bus.

Comment: This is a truly "classical" proof. In G.H. Hardy's *A Mathematician's Apology*, an extended rumination on the nature and beauty of pure mathematics, he gives just two examples of theorems: this theorem, and Euclid's proof of the infinitude of primes. As he says, this is inevitably a proof by contradiction (unlike Euclid's proof, which constructs new primes in a perfectly explicit way). The original statement is logically more complicated than what we actually prove in that it takes for granted that there is some *real* number $\sqrt{2}$ – characterized by being positive and having square equal to 2 – and then shows a "property" of this real number, namely it not being a fraction. But the essence of the matter is that a certain mathematical object *does not* exist – namely a rational number $\frac{a}{b}$ such that $(\frac{a}{b})^2 = 2$. This was the first "impossibility proof" in mathematics.

This is also one of the most historically important theorems in mathematics. History tells us that the result was discovered by Pythagoras, or at least someone in his school, and it was quite a shocking development (some sources say that the unnamed discoverer was fêted, others that he was cast into the sea). It caused Greek mathematicians to believe that geometric reasoning was more reliable than numerical, or quantitative reasoning, so that geometry became extremely well-developed in Greek mathematics at the expense of algebra.

Can we prove that $\sqrt{3}$ is irrational in the same way(s)? The Euclid's Lemma argument gives the irrationality of $\sqrt{p}$ for any prime p: write $\sqrt{p} = \frac{a}{b}$ in lowest terms, square and simplify to get $pb^2 = a^2$; then $p|a^2$ so $p|a$, so $a = pA$, and then substituting we get $pb^2 = p^2A^2$, $b^2 = pA^2$, so $p \mid b^2$ and finally $p \mid b$: contradiction.

It is interesting to notice that even without Euclid's Lemma we can prove the result "by hand" for any fixed prime p. For instance, with $p = 3$ we would like to prove: $3 \mid a^2 \implies 3 \mid a$. The contrapositive is that if a is not divisible by 3, neither is a^2. Since any number which is not divisible by 3 is of the form $3k + 1$ or $3k + 2$, we need only calculate:

$$(3k + 1)^2 = 9k^2 + 6k + 1 = 3(3k^2 + 2k) + 1,$$
$$(3k + 2)^2 = 9k^2 + 12k + 4 = 3(3k^2 + 4k + 1) + 1,$$

so in neither case did we get, upon squaring, a multiple of three. For any prime p, then, we can show $p|a^2 \implies p|a$ "by hand" by squaring each of the expressions $pk + i$, $0 < i < p$ and checking that we never get a multiple of p.

One can also look at this key step as a property of the ring Z_p of integers modulo p: if $0 \neq a \in Z_p$ then $0 \neq a^2 \in Z_p$. But – aha! – this is just saying that we don't want any nonzero elements in our ring Z_p which square to 0, so it will be true when Z_p is *reduced* (remember, this means that there are no nilpotent elements). When p is prime Z_p is an integral domain (even a field) so there are not even any zero divisors, but referring back to the algebra handout we proved more than this: for any n, Z_n is reduced iff n is squarefree. Thus, although the full strength of $p \mid ab \implies p \mid a$ or $p \mid b$ holds *only* for primes, the special case $p \mid a^2 \implies p \mid a$ is true not just for primes but for any squarefree integer p. (Stop and think about this for a moment; you can see it directly.) Thus the same argument in fact gives:

PROPOSITION 1.22. *For any squarefree integer $n > 1$, $\sqrt{n}$ is irrational.*

What about the case of general n? Well, of course $\sqrt{n^2}$ is not only rational but is an integer, namely n. Moreover, an arbitrary positive integer n can be factored to

get one of these two limiting cases: namely, any n can be uniquely decomposed as

$$n = sN^2,$$

where s is squarefree. (Prove it!) Since $\sqrt{sN^2} = N\sqrt{s}$, we have that $\sqrt{n}$ is rational iff $\sqrt{s}$ is rational; by the above result, this only occurs if $s = 1$. Thus:

THEOREM 1.23. *For $n \in \mathbb{Z}^+$, $\sqrt{n}$ is rational iff $n = N^2$ is a perfect square.*

Another way of stating this result is that $\sqrt{n}$ is either an integer or is irrational.

What about cube roots and so forth? We can prove that $\sqrt[3]{2}$ is irrational using a similar argument: suppose $\sqrt[3]{2} = \frac{a}{b}$, with $\gcd(a, b) = 1$. Then we get

$$2b^3 = a^3,$$

so $2 \mid a^3$, thus $2 \mid a$. Put $a = 2A$, so $b^3 = 2^2A^3$ and $2 \mid b^3$. Thus $2 \mid b$: contradiction.

Any integer can be written as the product of a cube-free integer[15] and a perfect cube; with this one can prove that the $\sqrt[3]{n}$ is irrational unless $n = N^3$. For the sake of variety, we prove the general result in a different way.

THEOREM 1.24. *Let $k > 2$ be a positive integer. Then $\sqrt[k]{n}$ is irrational unless $n = N^k$ is a perfect kth power.*

PROOF. Suppose n is not a perfect kth power. Then there is a prime $p \mid n$ such that $\mathrm{ord}_p(n)$ is not divisible by k. We use this prime to get a contradiction:

$$\frac{a^k}{b^k} = n, \; a^k = nb^k.$$

Take ord_p of both sides:

$$k \, \mathrm{ord}_p(a) = \mathrm{ord}_p(a^k) = \mathrm{ord}_p(nb^k) = k \, \mathrm{ord}_p(b) + \mathrm{ord}_p(n),$$

so $\mathrm{ord}_p(n) = k(\mathrm{ord}_p(a) - \mathrm{ord}_p(b))$ and $k \mid \mathrm{ord}_p(n)$: contradiction. □

From a more algebraic perspective, there is yet a further generalization to be made. A complex number α is an **algebraic number** if there exists a polynomial

$$P(t) = a_n t^n + \ldots + a_1 t + a_0$$

with $a_i \in \mathbb{Z}$, $a_n \neq 0$, such that $P(\alpha) = 0$. Similarly, α is an **algebraic integer** if there exists such a polynomial P with $a_n = 1$ (a **monic polynomial**). We write $\overline{\mathbb{Q}}$ for the set of algebraic numbers and $\overline{\mathbb{Z}}$ for the set of algebraic integers.

EXAMPLE 1.25. *$\alpha = \frac{1}{2} \in \overline{\mathbb{Q}}$ because α satisfies the polynomial $2t - 1$; $\beta = \sqrt[5]{2} \in \overline{\mathbb{Z}}$ because β satisfies the polynomial $t^5 - 2$. However there are also algebraic integers that "do not look like integers" in some naive sense: e.g. the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ satisfies the polynomial $t^2 - t - 1$ so is an algebraic integer.*

THEOREM 1.26. *If $\alpha \in \mathbb{Q} \cap \overline{\mathbb{Z}}$, then $\alpha \in \mathbb{Z}$.*

PROOF. Let $\alpha = \frac{a}{b}$ with $\gcd(a, b) = 1$; suppose α satisfies a monic polynomial:

$$\left(\frac{a}{b}\right)^n + c_{n-1}\left(\frac{a}{b}\right)^{n-1} + \ldots + c_1\left(\frac{a}{b}\right) + c_0 = 0, \; c_i \in \mathbb{Z}.$$

[15] I.e., an integer n with $\mathrm{ord}_p(n) \leq 2$ for all primes p.

We can clear denominators by multiplying through by b^n to get

$$a^n + bc_{n-1} \cdot a^{n-1} + \ldots + b^{n-1}c_1 \cdot a + b^n c_0 = 0,$$

or

$$a^n = b\left(-c_{n-1} \cdot a^{n-1} - \ldots - b^{n-2}c_1 \cdot a - b^{n-1}c_0\right). \tag{7}$$

If $b > 1$, then some prime p divides b and then, since p divides the right hand side of (34), it must divide the left hand side: $p \mid a^n$, so $p \mid a$. But, as usual, this contradicts the fact that a and b were chosen to be relatively prime. □

We can deduce Theorem 4 from Theorem 5 by noticing that for any k and n, $\sqrt[k]{n}$ is a root of the polynomial $t^k - n$ so lies in $\overline{\mathbb{Z}}$. On the other hand, evidently $\sqrt[k]{n}$ is an integer iff n is a perfect kth power, so when n is not a perfect kth power, $\sqrt[k]{n} \in \overline{\mathbb{Z}} \setminus \mathbb{Z}$, so by Theorem 1.26, $\sqrt[k]{n} \notin \mathbb{Q}$.

In fact Theorem 1.26 is a special case of a familiar result from high school algebra.

THEOREM 1.27. *(Rational Roots Theorem) If*

$$P(x) = a_n X + \ldots + a_1 x + a_0$$

is a polynomial with integral coefficients, then the only possible rational roots are those of the form $\pm\frac{c}{d}$, where $c \mid a_0$, $d \mid a_n$.

We leave the proof as an exercise. (It is similar to that of Theorem 1.26.)

6. Primitive Roots

Let N be a positive integer. An integer g is said to be a **primitive root** modulo N if every element x of $(\mathbb{Z}/N\mathbb{Z})^\times$ is of the form g^i for some positive integer i. Equivalently, the finite group $(\mathbb{Z}/N\mathbb{Z})^\times$ is cyclic and $g \pmod N$ is a generator.

We'd like to find primitive roots mod N, if possible. There are really two problems:

QUESTION 1. *For which N does there exist a primitive root modulo N?*

QUESTION 2. *Assuming there does exist a primitive root modulo N, how do we find one? How do we find all of them?*

We can and shall give a complete answer to Question 1. We already know that the group of units of a finite field is finite, and we know that $\mathbb{Z}/N\mathbb{Z}$ is a field if (and only if) N is prime. Thus primitive roots exist modulo N when N is prime.

When N is not prime we might as well ask a more general question: what is the structure of the unit group $(\mathbb{Z}/N\mathbb{Z})^\times$? From our work on the Chinese Remainder theorem, we know that if $N = p_1^{a_1} \cdots p_r^{a_r}$, there is an isomorphism of unit groups

$$(\mathbb{Z}/N\mathbb{Z})^\times = \mathbb{Z}/(p_1^{a_1} \cdots p_r^{a_r}\mathbb{Z})^\times \cong \prod_{i=1}^{r} (\mathbb{Z}/p_i^{a_i}\mathbb{Z})^\times.$$

Thus it is enough to figure out the group structure when $N = p^a$ is a prime power.

THEOREM 1.28. *The finite abelian group $(\mathbb{Z}/p^a\mathbb{Z})^\times$ is cyclic whenever p is an **odd** prime, or when $p = 2$ and a is 1 or 2. For $a \geq 3$, we have*

$$(\mathbb{Z}/2^a\mathbb{Z})^\times \cong Z_2 \times Z_{2^{a-2}}.$$

Before proving Theorem 1.28, let us nail down the answer it gives to Question 1.

COROLLARY 1.29. *Primitive roots exist modulo N in precisely the following cases:*
(i) $N = 1$, 2 or 4.
(ii) $N = p^a$ is an odd prime power.
(iii) $N = 2p^a$ is twice an odd prime power.

PROOF. Theorem 1.28 gives primitive roots in cases (i) and (ii). If p is odd, then

$$(\mathbb{Z}/2p^a\mathbb{Z})^\times \cong (\mathbb{Z}/2\mathbb{Z})^\times \times (\mathbb{Z}/p^a\mathbb{Z})^\times \cong (\mathbb{Z}/p^a\mathbb{Z})^\times$$

since $(\mathbb{Z}/2\mathbb{Z})^\times$ is the trivial group. Conversely, if N is not of the form (i), (ii) or (iii) then N is divisible either by 8 or by two distinct odd primes p and q. In the first case, write $N = 2^a \cdot M$ with $(2, M) = 1$ and $a \geq 3$. Then

$$(\mathbb{Z}/N\mathbb{Z})^\times \cong (\mathbb{Z}/2^a\mathbb{Z})^\times \times (\mathbb{Z}/M\mathbb{Z})^\times,$$

and $(\mathbb{Z}/N\mathbb{Z})^\times$, having the noncylic subgroup $(\mathbb{Z}/2^a\mathbb{Z})^\times$, cannot itself be cyclic [Handout A2.5, Corollary 6]. In the second case write $N = p^a q^b M$; then

$$(\mathbb{Z}/N\mathbb{Z})^\times \cong (\mathbb{Z}/p^a\mathbb{Z})^\times \times (\mathbb{Z}/q^b\mathbb{Z})^\times \times (\mathbb{Z}/M\mathbb{Z})^\times.$$

Both $(\mathbb{Z}/p^a\mathbb{Z})^\times$ and $(\mathbb{Z}/q^a\mathbb{Z})^\times$ have even order, hence their orders are not relatively prime and the product group cannot be cyclic [Handout A2.5, Corollary 10]. □

Proof of Theorem 1.28: The idea – for odd p – is as follows: if g is a primitive root mod p, then [Handout A2.5, Corollary 2] the order of g mod p^a is divisible by $p-1$, hence of the form $p^k \cdot (p-1)$ for some $k \leq a-1$. Therefore $g' = g^{p^k}$ has order $p-1$ [Handout A2.5, Proposition 7]. We claim $z = 1+p$ has order p^{a-1}; since $\gcd(p^{a-1}, p-1) = 1$, $g'z$ has order $p^{a-1}(p-1)$ [Handout A2.5, Example 4].

LEMMA 1.30. *Let p be an odd prime and $z \in \mathbb{Z}$, $z \equiv 1 \pmod{p}$.*
a) We have $\mathrm{ord}_p(z^p - 1) = \mathrm{ord}_p(z-1) + 1$.
b) For all $k \in \mathbb{Z}^+$, $\mathrm{ord}_p(z^{p^k} - 1) = \mathrm{ord}_p(z-1) + k$.

PROOF. Write $z = 1 + xp$ for some $x \in \mathbb{Z}$, so $\mathrm{ord}_p(z-1) = 1 + \mathrm{ord}_p(x)$. Then

$$(8)\quad z^p - 1 = (1+xp)^p - 1 = \binom{p}{1}(xp) + \binom{p}{2}(xp)^2 + \ldots + \binom{p}{p-1}(xp)^{p-1} + (xp)^p.$$

For the first term on the right hand side of (8), we have

$$\mathrm{ord}_p(\binom{p}{1}xp) = 2 + \mathrm{ord}_p(x) = \mathrm{ord}_p(z-1) + 1.$$

The remaining terms have larger p-orders, so the p-order of $z^p - 1$ is $\mathrm{ord}_p(z-1)+1$, whence part a). Since $z^{p^k} - 1 = (z^{p^{k-1}})^p - 1$, part b) follows by induction. □

Applying Lemma 1.30 to $z = 1 + p$ gives $\mathrm{ord}_p(z^{p^{k-1}} - 1) = k$ for all $k \in \mathbb{Z}^+$. So $z^{p^{a-2}} \neq 1 \pmod{p^a}$ and $z^{p^{a-1}} \equiv 1 \pmod{p^a}$: z has order exactly p^{a-1} in $(\mathbb{Z}/p^a\mathbb{Z})^\times$. Therefore, with notation as above, $g'z$ has order $p^{a-1}(p-1) = \#(\mathbb{Z}/p^a\mathbb{Z})^\times$, so is a primitive root mod p^a.

Now for $p = 2$. Note that $(\mathbb{Z}/2\mathbb{Z})^\times$ and $(\mathbb{Z}/4\mathbb{Z})^\times$ have orders 1 and 2 respectively so are certainly cyclic, and we may take $a \geq 3$. We claim that the subgroup

of $(\mathbb{Z}/2^a\mathbb{Z})^\times$ generated by 5 has order 2^{a-2} and is disjoint from the subgroup generated by -1, of order 2. It follows that the group is isomorphic to $Z_2 \times Z_{2^{a-2}}$.

When $p = 2$, Lemma 1.30 breaks down because the right hand side of (8) becomes just $4x + 4x^2 = 4x(x + 1)$, whose 2-order is at least $3 + \mathrm{ord}_2(x)$ if x is odd. So instead we take x even. In fact we may just take $x = 2$, so $z = 1 + 2x = 5$,

$$\mathrm{ord}_2(z^2 - 1) = \mathrm{ord}_2(z - 1) + \mathrm{ord}_2(z + 1) = \mathrm{ord}_2(z - 1) + \mathrm{ord}_2(6) = \mathrm{ord}_2(z - 1) + 1.$$

Again, inductively, we get

$$\mathrm{ord}_2(z^{2^k} - 1) = \mathrm{ord}_2(z - 1) + k,$$

or $\mathrm{ord}_2(5^{2^k} - 1) = k + 2$. Thus for $a \geq 2$, 5 has order 2^{a-2} in $(\mathbb{Z}/2^a\mathbb{Z})^\times$. Moreover $5^k + 1 \equiv 2 \pmod 4$ for all k, so $5^k \neq -1 \pmod{2^a}$, so the subgroups generated by the classes of 5 and of -1 are disjoint. This completes the proof of Theorem 1.28.

Question 2 remains: when there is a primitive root, then $(\mathbb{Z}N\mathbb{Z})^\times$ is a cyclic group, so has $\varphi(n)$ generators, where n is its order. Since the order of $(\mathbb{Z}/N\mathbb{Z})^\times$ is $n = \varphi(N)$, if there is one primitive root there are in fact exactly $\varphi(\varphi(N))$ of them, which is interesting. When $N = p$ is a prime, we get that there are $\varphi(p - 1)$ primitive roots. But how many is that?? We will turn to questions like this shortly.

Suppose now that $N = p$ is prime, so we know that there are a fair number of primitive roots modulo p, but how do we find one? This is a much deeper question. Suppose for instance we ask whether 2 is a primitive root modulo p. Well, it depends on p. Among odd primes less than 100, 2 **is** a primitive root modulo

3, 5, 11, 13, 19, 29, 37, 53, 59, 61, 67, 83

and *is not* a primitive root modulo

7, 17, 23, 31, 41, 43, 47, 71, 73, 79, 89, 97.

Extending the list, one finds that the "chance" that 2 is a primitive root modulo p seems to dip below $\frac{1}{2}$ and approach a number closer to 37%. E. Artin conjectured that for any prime number a, a is a primitive root modulo $(100C)\%$ of the primes, with

$$C = \prod_p \left(1 - \frac{1}{p(p-1)}\right) = 0.3739558136\ldots,$$

and in particular that a is a primitive root modulo infinitely many primes.

Work of Gupta-Murty[16] [**GM84**] and Heath-Brown[17] [**HB86**], shows that there are at most two "bad" prime numbers a such that a is a primitive root modulo only finitely many primes p. So, for instance, if 2 is not a primitive root modulo infinitely many primes and 3 is not either, than we can be sure that 5 is a primitive root modulo infinitely many primes!

[16]Two people: Rajiv Gupta and M.(aruti) Ram Murty.

[17]One person: D.(avid) Roger Heath-Brown.

There are further concrete questions of great interest: for instance, what can be said about the smallest primitive root mod p? Or, suppose we are given p and want to find a primitive root of p very quickly: what do we do? An extremely large literature exists on such matters.

CHAPTER 2

Pythagorean Triples

1. Parameterization of Pythagorean Triples

1.1. Introduction to Pythagorean triples.

By a **Pythagorean triple** we mean an ordered triple $(x, y, z) \in \mathbb{Z}^3$ such that

$$x^2 + y^2 = z^2.$$

The name comes from elementary geometry: if a right triangle has leg lengths x and y and hypotenuse length z, then $x^2 + y^2 = z^2$. Of course here x, y, z are positive real numbers. For most integer values of x and y, the integer $x^2 + y^2$ will not be a perfect square, so the positive real number $\sqrt{x^2 + y^2}$ will be irrational: e.g. $x = y = 1 \implies z = \sqrt{2}$. However, a few integer solutions to $x^2 + y^2 = z^2$ are familiar from high school algebra (and the SATs): e.g. $(3, 4, 5)$, $(5, 12, 13)$.

Remark: As soon as we have one solution, like $(3, 4, 5)$, we can find infinitely many more, however in a somewhat unsatisfying way. Namely, if (x, y, z) is a Pythagorean triple and a is any integer, then also (ax, ay, az) is a Pythagorean triple:

$$(ax)^2 + (ay)^2 = a^2(x^2 + y^2) = a^2 z^2 = (az)^2.$$

This property of invariance under scaling is a characteristic feature of solutions $(x_1, \ldots, x_n)$ to **homogeneous polynomials** $P(t_1, \ldots, t_n)$ in n-variables. We recall what this means: a monomial is an expression of the form $ct_1^{a_1} \cdots t_n^{a_n}$, and the degree of the monomial is defined to be $a_1 + \ldots + +a_n$, i.e., the sum of the exponents. A polynomial is said to be **homogeneous of degree d** if each of its monomial terms has degree d, and simply **homogeneous** if it is homogeneous of some degree d. For instance, the polynomial $P(x, y, z) = x^2 + y^2 - z^2$ is homogeneous of degree 2, and indeed for any N the **Fermat polynomial**

$$P_N(x, y, z) = x^N + y^N - z^N$$

is homogeneous of degree N. Moreover, every (nonconstant) homogeneous polynomial $P(t_1, \ldots, t_n)$ has zero constant term, hence $P(0, \ldots, 0) = 0$. So $(0, \ldots, 0)$ is a solution to any homogeneous polynomial, called the **trivial solution**.

Coming back to Pythagorean triples, these considerations show that for all $a \in \mathbb{Z}$, $(3a, 4a, 5a)$ is a Pythagorean triple (again, familiar to anyone who has studied for the SATs). For many purposes it is convenient to regard these rescaled solutions as being equivalent to each other. To this end we define a Pythagorean triple (a, b, c) to be **primitive** if $\gcd(a, b, c) = 1$. Then every nontrivial triple (a, b, c) is a positive integer multiple of a unique primitive triple, namely $(\frac{a}{d}, \frac{b}{d}, \frac{c}{d})$ where $d = \gcd(a, b, c)$.

Our goal is to find all primitive Pythagorean triples. There are many ways to do so. We prefer the following method, both for its simplicity and because it motivates the study of not just integral but **rational solutions** of polynomial equations.

Namely, consider the algebraic curve $x^2 + y^2 = 1$ in $\mathbb{R}^2$: i.e., the unit circle. Why? Well, suppose (a, b, c) is a nontrivial Pythagorean triple, so $a^2 + b^2 = c^2$ with $c \neq 0$ (if $c = 0$, then $a^2 + b^2 = 0 \implies a = b = 0$). So we may divide through by c, getting

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1.$$

Thus $(\frac{a}{c}, \frac{b}{c})$ is a rational point on the unit circle. Moreover, the process can be essentially reversed: suppose that $(r, s) \in \mathbb{Q}^2$ is such that $r^2 + s^2 = 1$. Then, writing $r = \frac{a}{c}$ and $s = \frac{b}{d}$ (so $cd \neq 0$), we have

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{d}\right)^2 = 1.$$

Multiplying through by $(cd)^2$, we get

$$(da)^2 + (bc)^2 = (bd)^2,$$

so that (da, bc, bd) is a nontrivial Pythagorean triple. If we start with a primitive Pythagorean triple (a, b, c), pass to the rational solution $(\frac{a}{c}, \frac{b}{c})$ and then clear denominators using the above formula, we get (ca, cb, c^2). This is not the primitive triple that we started with, but it is simply a rescaling: no big deal. At the end we will find the correct scaling that gives primitive triples on the nose.

1.2. Rational parameterization of the unit circle.

Fix any one rational point $P_\bullet = (x_\bullet, y_\bullet)$ on the unit circle. The argument that we are about to make works for any choice of $P_\bullet$ – e.g. $(\frac{3}{5}, \frac{4}{5})$ – but let me pass along the wisdom of hindsight: the computations will be especially simple and clean if we take $P_\bullet = (-1, 0)$. So let us do so.

Now suppose $P = (x_P, y_P)$ is any other rational point on the unit circle. Then there is a unique line ℓ joining $P_\bullet$ to P, which of course has rational coefficients:

$$\ell : y - y_P = \frac{y_P - y_\bullet}{x_P - x_\bullet}(x - x_P).$$

In particular, the slope of this line

$$m_P = \frac{y_P - y_\bullet}{x_P - x_\bullet},$$

is a rational number. This already places a limitation on the rational solutions, since "most" lines passing through the fixed point $P_\bullet$ have irrational slope. More interesting is the converse: for any $m \in \mathbb{Q}$, let

$$\ell_m : y = (y - y_\bullet) = m(x - x_\bullet) = m(x + 1),$$

be the line passing through $P_\bullet = (-1, 0)$ with slope m. We claim that this line intersects the unit circle in precisely one additional point P_m, and that this point P_m also has rational coordinates. That is, we claim that the rational points on the unit circle are precisely the point $P_\bullet = (-1, 0)$ together with the set of points P_m as m ranges through the rational numbers.

Why is this so? With a bit of thought, we can argue for this "in advance". Briefly, we plug the linear equation ℓ_m into the quadratic $x^2 + y^2 = 1$ thereby getting a quadratic equation in x with rational coefficients. Because we know that this equation has at least one rational solution – namely -1, the coordinate of $P_\bullet$ – the other solution must be rational as well, as follows from contemplation of the quadratic formula. On the other hand, such forethought is not really necessary in this case, because we want to find the solutions explicitly anyway. In other words, let's do it!

We have the system of equations

$$x^2 + y^2 = 1 \tag{9}$$

$$y = m(x + 1). \tag{10}$$

Substituting (40) into (34) gives

$$x^2 + m^2(x + 1)^2 = 1,$$

or

$$(1 + m^2)x^2 + 2m^2x + m^2 - 1 = 0.$$

Applying the quadratic formula, we get

$$x = \frac{-2m^2 \pm \sqrt{4m^4 - 4(1 + m^2)(m^2 - 1)}}{2(1 + m^2)}.$$

Under the radical sign we have

$$4m^4 - 4(m^2 + 1)(m^2 - 1) = 4(m^4 - (m^4 - 1)) = 4,$$

so that "luckily"[1] $\sqrt{4m^2 - 4(1 + m^2)(m^2 - 1)} = 2$, and

$$x = \frac{-2m^2 \pm 2}{2(1 + m^2)} = \frac{-m^2 \pm 1}{1 + m^2}.$$

Notice that by taking the minus sign, we get the solution $x = \frac{-m^2-1}{1+m^2} = -1$. That's great, because -1 is the x-coordinate of $P_\bullet$, so that it had better be a solution. The other solution is the one we really want:

$$x_m = \frac{1 - m^2}{1 + m^2},$$

and then we get

$$y_m = m(1 + x_m) = m\left(1 + \frac{1 - m^2}{1 + m^2}\right) = \frac{2m}{m^2 + 1},$$

so that finally

$$P_m = \left(\frac{1 - m^2}{1 + m^2}, \frac{2m}{1 + m^2}\right).$$

[1]Not really, of course: see the last paragraph above.

This is exactly what we wanted. Before returning to the problem of Pythagorean triples, however, let us make one further observation:

$$\lim_{m\to\pm\infty} P_m = \left(\lim_{m\to\pm\infty} \frac{1-m^2}{1+m^2}, \lim_{m\to\pm\infty} \frac{2m}{1+m^2}\right) = (-1,0) = P_\bullet.$$

The geometric interpretation of this is simple: the tangent line to the unit circle at $(-1,0)$ is vertical, so as the slope of the line ℓ_m approaches either $+\infty$ or $-\infty$, the second intersection point P_m approaches $P_\bullet$ and the secant lines approach the tangent line. So in fact it is true that the rational points on the unit circle correspond precisely to the set of all rational lines through $P_\bullet$: here we get $P_\bullet$ itself as the double intersection point of the tangent line. Thus, instead of $P_\bullet$, a more appropriate name would be P_∞, although we do not insist on this in the sequel.

1.3. Scaling to get primitive solutions.

We wish to explicitly write down all primitive Pythagorean triples (a,b,c). As above, this is accomplished up to scaling by clearing denominators in the general rational solution $P_m = (x_m, y_m)$. Namely, put $m = \frac{u}{v}$ with $\gcd(u,v) = 1$, so

$$P_m = \left(\frac{1-u^2/v^2}{1+u^2/v^2}, \frac{2u/v}{1+u^2/v^2}\right) = \left(\frac{v^2-u^2}{v^2+u^2}, \frac{2uv}{v^2+u^2}\right).$$

Thus, multiplying through by v^2+u^2, we get a family of integral solutions

$$(v^2-u^2, 2uv, v^2+u^2).$$

Are these solutions primitive? In other words, is $\gcd(v^2-u^2, 2uv, v^2+u^2) = 1$?

Suppose that an odd prime p divides v^2-u^2 and v^2+u^2. Then p also divides $(v^2-u^2)+(v^2+u^2) = 2v^2$ and $(v^2+u^2)-(v^2-u^2) = 2u^2$. Since p is odd, we get $p \mid u^2$ and $p \mid v^2$ which implies $p \mid u$ and $p \mid v$, contradiction. Similarly, if $4 \mid v^2-u^2$ and $4 \mid v^2+u^2$, then $4 \mid 2v^2$ and $4 \mid 2u^2$, so $2 \mid v^2$ and $2 \mid u^2$, so 2 divides both u and v. Thus $\gcd(v^2-u^2, 2uv, v^2+u^2)$ is either 1 or 2.

Case 1: v and u have opposite parity. Then v^2-u^2 is odd, so the gcd is 1. Notice that in this case, the first coordinate v^2-u^2 is odd and the second coordinate $2uv$ is even, so this can't be the complete list of all primitive Pythagorean triples: that set is symmetric under interchanging x and y!
Case 2: u and v are both odd. Then $v^2-u^2, 2uv, v^2+u^2$ are all even, so the gcd is 2. In this case $(\frac{v^2-u^2}{2}, uv, \frac{v^2+u^2}{2})$ is the primitive integral solution we seek.

This is the answer,[2] but let's touch it up a bit. If $x = 2k+1$ is odd, then $x^2 = 4k^2+4k+1 \equiv 1 \pmod 4$. Thus, if u and v are both odd, not only is v^2-u^2 even, it is congruent to $v^2-u^2 \equiv 1-1 = 0 \pmod 4$, so $\frac{v^2-u^2}{2}$ is even and uv is odd. Thus all the primitive triples arising in Case 2 are obtained by switching the first and second coordinates of a primitive triple in Case 1. To sum up:

THEOREM 2.1. *(Classification of Pythagorean Triples)*
a) The rational solutions to $x^2+y^2=1$ are $\{(-1,0) \cup \left(\frac{1-m^2}{1+m^2}, \frac{2m}{1+m^2}\right) \mid m \in \mathbb{Q}\}$.
*b) $(0,0,0)$ is a Pythagorean triple, called **trivial**. Every nontrivial Pythagorean*

[2]Note that there is no Case 3: we assumed $\gcd(u,v) = 1$, so they can't both be even.

triple is of the form (da, db, dc) *for some* $d \in \mathbb{Z}^+$*, where* (a, b, c) *is a Pythagorean triple with* $\gcd(a, b, c) = 1$*, called* ***primitive****.*
c) In every primitive Pythagorean triple (a, b, c)*, exactly one of* a *and* b *are even integers. Every primitive triple with* a *odd is of the form* $(v^2 - u^2, 2uv, v^2 + u^2)$ *where* $u, v \in \mathbb{Z}$ *are relatively prime integers of opposite parity. Conversely, all such pairs* u, v *yield a primitive Pythagorean triple with first coordinate odd.*

2. An Application: Fermat's Last Theorem for $N = 4$

In this section we will prove Fermat's Last Theorem for $N = 4$. In fact, following Fermat, we establish something stronger, from which FLT(4) immediately follows.

THEOREM 2.2. *(Fermat)* $X^4+Y^4 = Z^2$ *has no solutions with* $X, Y, Z \in \mathbb{Z}\backslash\{0\}$.

PROOF. Step 1: Let (x, y, z) be a positive integral solution to $X^4 + Y^4 = Z^2$. We claim there is a positive integral solution (x', y', z') with $\gcd(x', y') = 1$ and $z' \leq z$. Indeed, if x and y are not relatively prime, they are both divisible by some prime number p. Then $p^4 \mid X^4 + Y^4 = Z^2$, so $p^2 \mid Z$. Therefore $\frac{x}{p}, \frac{y}{p}, \frac{z}{p^2} \in \mathbb{Z}^+$ and

$$\left(\frac{x}{p}\right)^4 + \left(\frac{y}{p}\right)^4 = \frac{1}{p^4}(x^4 + y^4) = \frac{1}{p^4}(z^2) = \left(\frac{z}{p^2}\right)^2,$$

so $(\frac{x}{p}, \frac{y}{p}, \frac{z}{p^2})$ is another positive integral solution, with z-coordinate smaller than the one we started with. Therefore the process can be repeated, and since the z-coordinate gets strictly smaller each time, it must eventually terminate with a solution (x', y', z') as in the statement.
Step 2: Given a positive integral solution (x, y, z) to $X^4+Y^4 = Z^2$ with $\gcd(x, y) = 1$, we will produce another positive integral solution (u, v, w) with $w < z$.

First, we may assume without loss of generality that x is odd and y is even. They cannot both be even, since they are relatively prime; if instead x is even and y is odd, then we can switch x and y; so what we need to check is that x and y cannot both be odd. But then considering $x^4 + y^4 = z^2$ modulo 4, we find $2 \cong z^2 \pmod 4$, which is impossible: $0^2 \equiv 2^2 \equiv 0 \pmod 4$, $1^2 \equiv 3^2 \equiv 1 \pmod 4$.

Now we bring in our complete solution of Pythagorean triples: since $(x^2)^2 + (y^2)^2 = z^2$ and x and y are relatively prime, (x, y, z^2) is a primitive Pythagorean triple with first coordinate odd. Therefore by Theorem 2.1 there exist relatively prime integers m and n of opposite parity such that

$$\begin{aligned} x^2 &= m^2 - n^2 \\ y^2 &= 2mn \\ z &= m^2 + n^2. \end{aligned} \tag{11}$$

Now rewrite (27) as $n^2 + x^2 = m^2$. Since $\gcd(m, n) = 1$, this is again a primitive Pythagorean triple. Moreover, since x is odd, n must be even. So we can use our parameterization again (!!) to write

$$\begin{aligned} x &= r^2 - s^2, \\ n &= 2rs, \\ m &= r^2 + s^2, \end{aligned}$$

for coprime integers r, s of oposite parity. Now observe

$$m\left(\frac{n}{2}\right) = \frac{2mn}{4} = \frac{y^2}{4} = \left(\frac{y}{2}\right)^2.$$

Since m and $\frac{n}{2}$ are coprime integers whose product is a perfect square, they must both be perfect squares. Similarly,

$$rs = \frac{2rs}{2} = \frac{n}{2} = \square,$$

so r and s must both be squares. Let us put $r = u^2, s = v^2, m = w^2$, and substitute these quantities into $m = r^2 + s^2$ to get

$$u^4 + v^4 = w^2.$$

Here $w \geq 1$, so

$$w \leq w^4 < w^4 + n^2 = m^2 + n^2 = z,$$

so that as promised, we found a new positive integral solution (u, v, w) with $w < z$. Step 3: Steps 1 and 2 together lead to a contradiction, as follows: if we have any positive integral solution (x, y, z) to $X^4 + Y^4 = Z^2$, then by Step 1 we have one (x', y', z') with $z' \leq z$ with $\gcd(x', y') = 1$. Then by Step 2 we have another positive integral solution (u, v, w) with $w < z' \leq z$. Then by Step 1 we have another positive integral solution (u', v', w') with $w' \leq w < z' \leq z$ with $\gcd(u', v') = 1$, and then by Step 2 we get another solution whose final coordinate is strictly smaller than w. And so on. In other words, the assumption that there are any positive integer solutions at all leads to the construction of an infinite sequence of positive integer solutions (x_n, y_n, z_n) with $z_{n+1} < z_n$ for all n. But that's impossible: there are no infinite strictly decreasing sequences of positive integers. Contradiction! □

LEMMA 2.3. *Let A and B be coprime integers. Then:*
a) If A and B have the same parity, $\gcd(A + B, A - B) = 2$.
b) If A and B have opposite parity, $\gcd(A + B, A - B) = 1$.

Exercise: Prove Lemma 2.3.

Here is a second proof, communicated to us by Barry Powell.

PROOF. Seeking a contradiction, suppose the equation $X^4 + Y^4 = Z^2$ has solutions (x, y, z) with $z \neq 0$. Among all such solutions, choose one with z^2 *minimal*. For such a minimal solution we must have $\gcd(x, y) = 1$: indeed, if a prime p divided both x and y, then $p^4 \mid z^2$ so $p^2 \mid z$ and we may take $x = px'$, $y = py'$, $z = p^2 z'$ to get a solution (x', y', z') with $(z')^2 < z^2$. Moreover x and y must have opposite parity: being coprime they cannot both be even; if both were odd then reducing modulo 4 gives a contradiction. It is no loss of generality to assume that x is odd and y is even, and it follows that z is odd.

We claim that $\gcd(z+y^2, z-y^2) = 1$. Indeed, let $d = \gcd(z+y^2, z-y^2)$. Since y is even and z is odd, $z+y^2$ is odd, hence d is odd. Suppose p is an odd prime dividing d. Then $p \mid (z+y^2)+(z-y^2) = 2z$, so $p \mid z$; moreover, $p \mid (z+y^2)-(z-y^2) = 2y^2$, so $p \mid y$. Since $x^4 = z^2 - y^4$, it follows that $p \mid x$, contradicting $\gcd(x, y) = 1$.

By uniqueness of factorization there are coprime integers r and s such that

$$z - y^2 = r^4, \ z + y^2 = s^4. \tag{12}$$

So $(s^2+r^2)(s^2-r^2) = s^4 - r^4 = 2y^2$ with y even, hence r and s are both odd. Since s^2, r^2 are coprime integers of the same parity, by Lemma 2.3, $\gcd(s^2+r^2, s^2-r^2) = 2$. Since r, s are odd, so is $\frac{s^2+r^2}{2}$, so $\gcd(\frac{s^2+r^2}{2}, s^2-r^2) = 1$, and then by uniqueness of factorization there are coprime integers a, b with

$$r^2 + s^2 = 2b^2, \ (r+s)(r-s) = r^2 - s^2 = a^2.$$

Again by Lemma 2.3, $\gcd(r+s, r-s) = 2$, so $\gcd(\frac{r+s}{2}, \frac{r-s}{2}) = 1$ and thus there are coprime integers u, v with

$$s - r = 2u^2, s + r = 2v^2.$$

It follows that

$$4(u^4 + v^4) = (s-r)^2 + (s+r)^2 = 2(s^2 + r^2) = 4b^2,$$

and thus

$$u^4 + v^4 = b^2.$$

Since x is odd, hence nonzero, and $x^4 + y^4 = z^2$, $y^2 \leq y^4 < z^2$, so

$$2b^2 = s^2 + r^2 \leq (s^2 + r^2)(s^2 - r^2) = 2y^2 < 2z^2.$$

Note also that $b \neq 0$, since otherwise $u = v = 0$, contradicting the fact that they are coprime. Therefore we have found a solution (u, v, b) to $X^4 + Y^4 = Z^2$ with $0 < b^2 < z^2$, contradicting the minimality of z^2. □

COROLLARY 2.4. *$X^4 + Y^4 = Z^4$ has no solutions with $X, Y, Z \in \mathbb{Z} \setminus \{0\}$.*

PROOF. Suppose there are $x, y, z \in \mathbb{Z} \setminus \{0\}$ such that $x^4 + y^4 = z^4$. We may assume x, y, z are all positive. Then, since $Z^4 = (Z^2)^2$, the triple (x, y, z^2) is a positive integer solution to $X^4 + Y^4 = Z^2$, contradicting Theorem 2.2. □

The strategy of the above proof is known as **infinite descent**. Over the centuries it has been refined and developed, and the modern **theory of descent** is one of the mainstays of contemporary Diophantine geometry.

3. Rational Points on Conics

The method of drawing lines that we used to find all rational points on the unit circle has further applicability. Namely, we consider an arbitrary **conic curve**

$$aX^2 + bY^2 = cZ^2, \tag{13}$$

for $a, b, c \in \mathbb{Q} \setminus \{0\}$.

Remark: More generally, one calls a plane conic any curve given by an equation

$$aX^2 + bXY + cXZ + dY^2 + eYZ + fZ^2 = 0.$$

for $a, b, c, d, e, f \in \mathbb{Q}$, not all zero. But as one learns in linear algebra, by making a linear change of variables, new coordinates can be found in which the equation is diagonal, i.e., in the form (13), and one can easily relate integral/rational points on one curve to those on the other. So by considering only diagonalized conics, we are not losing out on any generality.

Now, as in the case $a = b = c = 1$, we have a bijective correspondence between primitive integral solutions to $aX^2 + bY^2 = cZ^2$ and rational points on

$$ax^2 + by^2 = c. \tag{14}$$

If we can find any one rational point $P_\bullet = (x_\bullet, y_\bullet)$ on (14) then our previous method works: taking the set of all lines through $P_\bullet$ with rational slope, together with the vertical line $x = x_\bullet$ and intersecting with the conic (14), we get all rational solutions.

In the exercises the reader is invited to try this in certain cases where there are

"obvious" rational solutions. For instance, if $a = c$ then an obvious rational solution is $(1, 0)$. The reader is asked to carry this out in a particular case – and also to investigate the structure of the primitive integral solutions – in the exercises.

But there need not be any rational solutions at all! An easy example of this is

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

where indeed there are clearly no $\mathbb{R}$-solutions. But this is not the only obstruction. Consider for instance

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

whose real solutions form a circle of radius $\frac{1}{\sqrt{3}}$. We claim that there are however no rational points on this circle. Equivalently, there are no integral solutions to $3X^2 + 3Y^2 = Z^2$ with $\gcd(x, y, z) = 1$. For suppose there is such a primitive integral solution (x, y, z). Then, since $3 \mid 3x^2 + 3y^2 = z^2$, we have $3 \mid z$. So we may put $z = 3z'$, getting $3x^2 + 3y^2 = 9(z')^2$, or

$$x^2 + y^2 = 3z^2.$$

Now reducing mod 3, we get

$$x^2 + y^2 \equiv 0 \pmod 3.$$

Since the squares mod 3 are 0 and 1, the only solution mod 3 is $x \equiv y \equiv 0 \pmod 3$, but this means $3 \mid x$, $3 \mid y$, so that the solution (x, y, z) is not primitive after all: 3 is a common divisor.

This argument can be made to go through with 3 replaced by any prime p with $p \equiv 3 \pmod 4$. Arguing as above, it suffices to show that the congruence $x^2 + y^2 = 0 \pmod p$ has only the zero solution. But if it has a solution with, say, $x \neq 0$, then x is a unit modulo p and then $\left(\frac{y}{x}\right)^2 = -1 \pmod p$. We will see later that for an odd prime p, the equation $a^2 \equiv -1 \pmod p$ has a solution iff $p \equiv 1 \pmod 4$.

In fact, for an odd prime $p \equiv 1 \pmod 4$, the curve

$$px^2 + py^2 = 1$$

does always have rational solutions, although this is certainly not obvious. Overall we need a method to decide whether the conic $aX^2 + bY^2 = cZ^2$ has any nontrivial integral solutions. This is provided by the following elegant theorem of Legendre.

THEOREM 2.5. *Let a, b, c be nonzero squarefree integers, relatively prime in pairs, and neither all positive nor all negative. Then*

$$ax^2 + by^2 + cz^2 = 0$$

has a solution in nonzero integers (x, y, z) iff all of the following hold:
(i) There exists $x \in \mathbb{Z}$ such that $-ab \equiv x^2 \pmod{|c|}$.
(ii) There exists $y \in \mathbb{Z}$ such that $-bc \equiv y^2 \pmod{|a|}$.
(iii) there exists $z \in \mathbb{Z}$ such that $-ca \equiv z^2 \pmod{|b|}$.

In particular, since we can compute all of the squares modulo any integer n by a direct, finite calculation, we can easily program a computer to determine whether or not the equation has any nonzero integer solutions. Once we know whether there are any integral solutions, we can search by brute force until we find one. The following result of Holzer puts an explicit upper bound on our search:

THEOREM 2.6. *If the equation $ax^2+by^2+cz^2 = 0$ has any solutions in nonzero integers, it has such a solution (x, y, z) with $|x| \leq \sqrt{bc}$, $|y| \leq \sqrt{ac}$, $|z| \leq \sqrt{ab}$.*

A short and elementary proof of Theorem 2.6 was given by L.J. Mordell [**Mo69**]. Mordell's proof was somewhat terse in places, leading to certain claims of a gap in his argument. In her final project for the 2009 course, Laura Nunley closely examined [**Mo69**] and found that it is complete and correct. A more discursive writeup of the argument appears in her 2010 UGA master's thesis [**Nu10**]. Nunley's thesis also contains a detailed treatment of a beautiful generalization of Theorem 2.6 due to Cochrane and Mitchell, which we will discuss later on.

Thus the study of homogeneous quadratic equations over $\mathbb{Z}$ (or, what comes to the same, over $\mathbb{Q}$) is admirably complete. The same cannot be said for polynomial equations of higher degree, as we will soon see.

CHAPTER 3

Quadratic Rings

1. Quadratic Fields and Quadratic Rings

Let D be a squarefree integer not equal to 0 or 1. Then $\sqrt{D}$ is irrational, and $\mathbb{Q}[\sqrt{D}]$, the subring of $\mathbb{C}$ obtained by adjoining $\sqrt{D}$ to $\mathbb{Q}$, is a field.

From an abstract algebraic perspective, an explanation for this can be given as follows: since $\sqrt{D}$ is irrational, the polynomial $t^2 - D$ is irreducible over $\mathbb{Q}$. Since the ring $\mathbb{Q}[t]$ is a PID, the irreducible element $t^2 - D$ generates a maximal ideal $(t^2 - D)$, so that the quotient $\mathbb{Q}[t]/(t^2 - D)$ is a field. Moreover, the map $\mathbb{Q}[\sqrt{D}] \to \mathbb{Q}[t]/(t^2 - D)$ which is the identity on $\mathbb{Q}$ and sends $\sqrt{D} \mapsto t$ is an isomorphism of rings, so $\mathbb{Q}[\sqrt{D}]$ is also a field. We may write $\mathbb{Q}[\sqrt{D}] = \{a + b\sqrt{D} \mid a, b \in \mathbb{Q}\}$, so that a basis for $\mathbb{Q}[\sqrt{D}]$ as a $\mathbb{Q}$-vector space is $1, \sqrt{D}$. In particular $\mathbb{Q}[\sqrt{D}]$ is two-dimensional as a $\mathbb{Q}$-vector space: we accordingly say it is a **quadratic field**.

It is also easy to check by hand that the ring $\mathbb{Q}[\sqrt{D}]$ is a field. For this and for many other things to come, the key identity is

$$(a + b\sqrt{D})(a - b\sqrt{D}) = a^2 - Db^2.$$

For rational numbers a and b which are not both zero, the rational number $a^2 - Db^2$ is also nonzero: equivalently there are no solutions to $D = \frac{a^2}{b^2}$, because $\sqrt{D}$ is irrational. It follows that – again, for a, b not both 0 – we have

$$\left(a + b\sqrt{D}\right) \cdot \left(\frac{a}{a^2 - Db^2} - \frac{b}{a^2 - Db^2}\sqrt{D}\right) = 1,$$

which gives a multiplicative inverse for $a + b\sqrt{D}$ in $\mathbb{Q}[\sqrt{D}]$.

We wish also to consider **quadratic rings**, certain integral domains whose fraction field is a quadratic field $\mathbb{Q}(\sqrt{D})$. Eventually we will want a more precise and inclusive definition, but for now we consider $\mathbb{Z}[\sqrt{D}] = \{a + b\sqrt{D} \mid a, b \in \mathbb{Z}\}$.[1]

2. Fermat's Two Squares Theorem

The rings $\mathbb{Z}[\sqrt{D}]$ occur naturally when we study Diophantine equations. E.g:

[1]This equality is a fact which is not difficult to check; it is *not* the definition of $\mathbb{Z}[\sqrt{D}]$. By way of comparison, we recommend that the reader check that the ring $\mathbb{Z}[\frac{\sqrt{D}}{2}]$ is *not* of the form $\mathbb{Z}\alpha + \mathbb{Z}\beta$ for any two fixed elements α, β of $\mathbb{Z}[\frac{\sqrt{D}}{2}]$. In fact its additive group is not finitely generated as an abelian group.

QUESTION 3. *Which prime numbers p can be expressed as a sum of two squares? More precisely, for for which prime numbers p are there integers x and y such that*

$$x^2 + y^2 = p? \tag{15}$$

Evidently 2 is a sum of squares: $1^2 + 1^2 = 2$. Henceforth we assume $p > 2$.

At the moment we have exactly one general technique[2] for studying Diophantine equations: congruences. So let's try to apply it here.

• If we reduce $x^2 + y^2 = p$ modulo 4, we get that p is a sum of two squares modulo 4. Since $0^2 \equiv 2^2 \equiv 0 \pmod 4$ and $1^2 \equiv 3^2 \equiv 1 \pmod 4$, the squares modulo 4 are $\{0, 1\}$, and thus – please stop and do the calculation yourself! – the sums of two squares modulo 4 are $\{0, 1, 2\}$. Especially, 3 is not a sum of two squares modulo 4. We deduce:

LEMMA 3.1. *If an odd prime p is a sum of two squares, then $p \equiv 1 \pmod 4$.*

• Suppose $x^2 + y^2 = p$. Reducing modulo p we get $x^2 + y^2 \equiv 0 \pmod p$. If $y \equiv 0 \pmod p$, then $x^2 \equiv 0 - y^2 \equiv 0 \pmod p$ and thus also $x \equiv 0 \pmod p$, so we may take $x = pX$, $y = pY$ to get

$$p = x^2 + y^2 = p^2(X^2 + Y^2)$$

and thus $p^2 \mid p$, a contradiction. So $y \in (\mathbb{Z}/p\mathbb{Z})^\times$ and we may divide through by y^2, getting

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

That is, a necessary condition for (34) to have solutions is that -1 be a square modulo p. It turns out though that this is not new information.

PROPOSITION 3.2. *Let p be an odd prime, and let $x \in U(p) = (\mathbb{Z}/p\mathbb{Z})^\times$. Then:*
a) x is a square in $U(p)$ iff $x^{\frac{p-1}{2}} = 1$.
b) In particular, -1 is a square modulo p iff $p \equiv 1 \pmod 4$.

PROOF. a) Write $U(p) = (\mathbb{Z}/p\mathbb{Z})^\times$, and consider the map

$$\Phi : U(p) \to U(p),\ x \mapsto x^{\frac{p-1}{2}}.$$

Let $U(p)^2 = \{x^2 \mid x \in U(p)\}$ be the subgroup of squares. The assertion of part a) is equivalent to: $\operatorname{Ker}\Phi = U(p)^2$. We now demonstrate this: since $x \mapsto x^2$ is a homomorphism from the group $U(p)$ of order $p-1$ to itself with kernel $\{\pm 1\}$ of order 2, we have $\#U(p)^2 = \frac{p-1}{2}$. Further, for $x \in U(p)$, $(x^2)^{\frac{p-1}{2}} = x^{p-1} = 1$ by Lagrange's Little Theorem, so $U(p)^2 \subset \operatorname{Ker}\Phi$. On the other hand, every $x \in \operatorname{Ker}\Phi$ satisfies $x^{\frac{p-1}{2}} - 1 = 0$, and a polynomial over a field cannot have more roots than its degree, so $\#\operatorname{Ker}\Phi \leq \frac{p-1}{2}$. Therefore $U(p)^2 = \operatorname{Ker}\Phi$.
b) Take $x = -1$ in part a): $(-1)^{\frac{p-1}{2}} \equiv 1 \pmod p$ iff $\frac{p-1}{2}$ is even iff $p \equiv 1 \pmod 4$. □

[2]By a "general technique", we mean a technique that can always be applied, not one that is guaranteed always to succeed. In that stronger sense there are provably no general technique for Diophantine equations!

Most of the point of the above proof is that it does not use the cyclicity of $U(p)$; however it still uses a good working familiarity with basic group theory.

EXERCISE 3.1. Use the cyclicity of $U(p)$ to give a quicker proof of Proposition 3.2.

As it happens, in order to determine which primes are a sum of 2 squares we only need half of the above result, and that half has a more elementary proof.

LEMMA 3.3. *(Fermat's Lemma) For a prime $p \equiv 1 \pmod 4$, there is an integer x such that $p \mid x^2 + 1$. Equivalently, -1 is a square modulo p.*

PROOF. A **reduced residue system modulo p** is a set S of $p-1$ integers such that the reduction $\overline{S}$ of S modulo p is precisely the set $(\mathbb{Z}/p\mathbb{Z})^\times$ of nonzero residues. By Wilson's Theorem, for any reduced residue system S, we have $\prod_{x \in S} x \equiv -1 \pmod p$. The most obvious choice of a reduced residue system modulo p is of course $\{1, \ldots, p-1\}$. To prove this result we will use the second most obvious choice, namely

$$S = \{\frac{-(p-1)}{2}, \frac{-(p-1)}{2} + 1, \ldots, -1, 1, \ldots, \frac{(p-1)}{2}\}.$$

Then

$$-1 \equiv \prod_{x \in S} x \equiv (-1)^{\frac{p-1}{2}} (\left(\frac{p-1}{2}\right)!)^2 \pmod p.$$

If $p \equiv 1 \pmod 4$, then $\frac{p-1}{2}$ is even, $(-1)^{\frac{p-1}{2}} = 1$, and -1 is a square modulo p. □

It follows from Fermat's Lemma that (15) has no $\mathbb{Z}$-solutions unless $p \equiv 1 \pmod 4$. What about the converse: if $p \equiv 1 \pmod 4$, must p be a sum of two squares?

By Fermat's Lemma, there is $x \in \mathbb{Z}$ such that $x^2 \equiv -1 \pmod p$, i.e., there exists $n \in \mathbb{Z}$ such that $pn = x^2 + 1$. Now factor the right hand side over $\mathbb{Z}[\sqrt{-1}]$:

$$pn = (x + \sqrt{-1})(x - \sqrt{-1}).$$

Suppose that p is prime as an element of $\mathbb{Z}[\sqrt{-1}]$. Then it satisfies Euclid's Lemma: if $p \mid \alpha\beta$, then $p \mid \alpha$ or $p \mid \beta$. Here, if p is prime in $\mathbb{Z}[\sqrt{-1}]$, then we get $p \mid x \pm i$. But this is absurd: what this means is that the quotient $\frac{x \pm i}{p} = \frac{x}{p} \pm \frac{1}{p} i$ is an element of $\mathbb{Z}[\sqrt{-1}]$, i.e., that both $\frac{x}{p}$ and $\frac{1}{p}$ are integers. But obviously $\frac{1}{p}$ is not an integer. Therefore p is not prime, so[3] there exists a nontrivial factorization

$$p = \alpha\beta, \tag{16}$$

where $\alpha = a + b\sqrt{-1}, \beta = c + d\sqrt{-1} \in \mathbb{Z}[\sqrt{-1}]$ are nonunit elements. Taking complex conjugates of the above equation, we get

$$\overline{p} = p = \overline{\alpha\beta} = \overline{\alpha\beta}. \tag{17}$$

Multiplying (40) and (27) we get

$$p^2 = (\alpha\overline{\alpha})(\beta\overline{\beta}) = (a^2 + b^2)(c^2 + d^2). \tag{18}$$

Now, since α and β are evidently nonzero, we have $a^2 + b^2,\ c^2 + d^2 > 0$. We claim that indeed $a^2 + b^2 \neq 1$ and $c^2 + d^2 \neq 1$. Indeed, if $a + b\sqrt{-1} \in \mathbb{Z}[\sqrt{-1}]$

[3]A gap occurs in the argument here. It has been deliberately inserted for pedagogical reasons. Please keep reading at least until the beginning of the next section!

with $a^2+b^2=1$, then its multiplicative inverse in $\mathbb{Q}[\sqrt{-1}]$ is $\frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}\sqrt{-1} = a - b\sqrt{-1}$ which again lies in $\mathbb{Z}[\sqrt{-1}]$. In other words, $a^2+b^2=1$ implies that $a+b\sqrt{-1}$ is a unit in $\mathbb{Z}[\sqrt{-1}]$, contrary to our assumption. Having ruled out that either $a^2+b^2=1$ or $c^2+d^2=1$, (28) now immediately implies

$$a^2+b^2=c^2+d^2=p.$$

But that is what we wanted: p is a sum of two squares! Thus we have (apparently...please read on!) proved the following theorem.

THEOREM 3.4. *(Fermat's Two Squares Theorem) A prime number p can be expressed as the sum of two integer squares if and only if $p=2$ or $p \equiv 1 \pmod 4$.*

3. Fermat's Two Squares Theorem Lost

The above proof of Theorem 3.4 was surprisingly quick and easy, especially compared to Fermat's original one: not having the notion of factorization in domains other than the integers, Fermat's uses (as is typical of him) a more intricate argument by descent. In fact, the way we have presented the above argument, it is **too easy**: there is a gap in the proof. The gap is rather subtle, and is analogous to a notorious mistake made by the early 19th century mathematician Lamé. Rather than expose it directly, let us try to squeeze more out of it and see what goes wrong.

Namely, instead of just working with $x^2+y^2=p$ and the corresponding quadratic ring $\mathbb{Z}[\sqrt{-1}]$, let us consider the equation

$$x^2 - Dy^2 = p, \tag{19}$$

where p is still a prime and D is a squarefree integer different from 0 or 1. We can mimic the above argument in two steps as follows:

Step 1: By reducing modulo p, we get exactly as before that the existence of an integral solution to (19) implies that D is a square modulo p.

Step 2: Conversely, assume that D is a square modulo p, i.e., there exists $x \in \mathbb{Z}$ such that $D \equiv x^2 \pmod p$. Again this leads means there exists $n \in \mathbb{Z}$ such that

$$pn = x^2 - D,$$

and thus leads to a factorization in $\mathbb{Z}[\sqrt{D}]$, namely

$$pn = (x+\sqrt{D})(x-\sqrt{D}).$$

Now if p were a prime element in $\mathbb{Z}[\sqrt{D}]$ it would satisfy Euclid's Lemma, and therefore since it divides the product $(x+\sqrt{D})(x-\sqrt{D})$, it must divide one of the factors: $p \mid x \pm \sqrt{D}$. But since $\frac{x}{p} \pm \frac{1}{p}\sqrt{D}$ is still not in $\mathbb{Z}[\sqrt{D}]$, this is absurd: p is not a prime element in $\mathbb{Z}[\sqrt{D}]$. So it factors nontrivially: $p=\alpha\beta$, for α, β nonunits in $\mathbb{Z}[\sqrt{D}]$. Let us now define, for any $\alpha = a+b\sqrt{D} \in \mathbb{Q}(\sqrt{D})$, $\overline{\alpha} = a - b\sqrt{D}$. When $-D<0$ this is the usual complex conjugation. When $-D>0$ it is the conjugate in the sense of high school algebra (and also in Galois theory). It is entirely straightforward to verify the identity

$$\overline{\alpha\beta} = \overline{\alpha}\,\overline{\beta}$$

in either case, so that again by taking conjugates we get also

$$p = \overline{p} = \overline{\alpha}\overline{\beta},$$

and multiplying the two equations we get

$$p^2 = (\alpha\overline{\alpha})(\beta\overline{\beta}) = (a^2 - Db^2)(c^2 - Db^2).$$

As above, if $a^2 - Db^2 = \pm 1$, then the inverse of α lies in $\mathbb{Z}[\sqrt{D}]$, so α is a unit in $\mathbb{Z}[\sqrt{D}]$, which we assumed it not to be. This time, the conclusion we get is

$$p = \pm(a^2 - Db^2).$$

In particular, if $D < 0$, the conclusion we get is that p is of the form $a^2 + |D|b^2$ if and only if $-D$ is a square mod p.

Unfortunately we can easily see that this conclusion is very often wrong. Namely, suppose $D \leq -3$ and take $p = 2$.[4] Then the condition that $-D$ is a square modulo 2 is quite vacuous: the only elements of $\mathbb{Z}/2\mathbb{Z}$ are $0 = 0^2$ and $1 = 1^2$, so every integer is congruent to a square modulo 2. Thus the above argument implies there are integers x and y such that

$$2 = x^2 + |D|y^2.$$

But this is absurd: if $y = 0$ it tells us that 2 is a square; whereas if $y \geq 1$, $x^2 + |D|y^2 \geq |D| \geq 3$. In other words, 2 is certainly *not* of the form $x^2 + |D|y^2$.

So what went wrong?!?

4. Fermat's Two Squares Theorem (and More!) Regained

It is time to come clean. We have been equivocating over the definition of a prime element in an integral domain. Recall that we did not actually define such a thing. Rather, we defined an **irreducible** element in R to be a nonzero nonunit f such that $f = xy$ implies x or y is a unit. Then in the integers we proved Euclid's Lemma: if an irreducible element f of $\mathbb{Z}$ divides ab, then either f divides a or f divides b. Of course this was not obvious: rather, it was all but equivalent to the fundamental theorem of arithmetic.

Let us now review how this is the case. By definition, a domain R is a unique factorization domain if it satisfies two properties: first, that every nonzero nonunit factor into irreducible elements, and second that this factorization be unique, up to ordering of factors and associate elements.

Suppose R is a UFD. We claim that if f is an irreducible element of R, and $f \mid ab$, then $f \mid a$ or $f \mid b$. We already proved this for $R = \mathbb{Z}$ and the general argument is the same: we leave to the reader the very important exercise of looking back at the argument for $R = \mathbb{Z}$ and adapting it to the context of R a UFD.

We define a **prime element** in a domain R to be a nonzero nonunit p which satisfies the conclusion of Euclid's Lemma: if $p \mid ab$, then $p \mid a$ or $p \mid b$.

[4]Taking $p = 2$ should always raise alarm bells in your head: it is often said that "2 is the oddest prime." In this case, please do look back the above argument to see that $p = 2$ was not – and did not need to be – excluded from consideration.

PROPOSITION 3.5. *Let R be an integral domain.*
a) Any prime element is irreducible.
b) R is a UFD if and only if it is a factorization domain in which every irreducible element is prime.

Because this is pure algebra, we prefer to not discuss the proof here. It is a good exercise for the reader. See also
`http://www.math.uga.edu/~pete/factorization.pdf`
for a careful treatment of the theory of factorization in integral domains.

We can now recast the results of the previous two sections as follows. First, the "proof" which we gave assumed that either p is a prime element of $\mathbb{Z}[\sqrt{D}]$ or that p is not an irreducible element: i.e., it factors as $p = \alpha\beta$ with α, β nonunit elements of $\mathbb{Z}[\sqrt{D}]$. What was missing was the third possibility: p is an irreducible element of $\mathbb{Z}[\sqrt{D}]$ which is not prime. Because of Proposition 3.5, this third possibility cannot occur if $\mathbb{Z}[\sqrt{D}]$ is a UFD, so what we actually proved was:

THEOREM 3.6. *Let D be a squarefree integer different from 0 and 1. We **assume** that the ring $\mathbb{Z}[\sqrt{D}]$ is a UFD. Then, for a prime number p, TFAE:*
(i) There exist $x, y \in \mathbb{Z}$ such that $p = |x^2 - Dy^2|$.
(ii) There exists $x \in \mathbb{Z}$ such that $D \equiv x^2 \pmod{p}$.

Moreover, simply by noticing that for $D < -2$ we had that D is a square mod 2 but 2 is not of the form $|x^2 - Dy^2|$, we also deduce:

COROLLARY 3.7. *For no $D < -2$ is the ring $\mathbb{Z}[\sqrt{D}]$ a UFD.*

To complete the proof of Theorem 3.4 it suffices to show that at least $\mathbb{Z}[\sqrt{-1}]$ is a UFD. Fortunately for us, it is. We show this in a way which is again a generalization of one of our proofs of the fundamental theorem of arithmetic: namely, by showing that we can perform, in a certain appropriate sense, division with remainder in $\mathbb{Z}[\sqrt{-1}]$. The formalism for this is developed in the next section.

4.1. Euclidean norms.

A norm $N : R \to \mathbb{N}$ is called **Euclidean** if it satisfies the following property: for all $a \in R, b \in R \setminus \{0\}$, there exist $q, r \in R$ such that $a = qb + r$ and $N(r) < N(b)$.

First let us see that this suffices: we claim that any domain R endowed with a Euclidean norm function is a principal ideal domain. Indeed, let I be any nonzero ideal of such a domain R, and let a be any element of minimal nonzero norm. We claim that in fact $I = (b) = \{rb \mid r \in R\}$. The proof is the same as for integers: suppose that $a \in I$ and apply the Euclidean property: there exist $q, r \in R$ such that $a = qb + r$ with $N(r) < N(b)$. But $r = a - qb$ and $a, b \in I$, so $r \in I$. If $r \neq 0$ then $N(r) \neq 0$ and we have found an element with nonzero norm smaller than $N(a)$, contradiction. So we must have $r = 0$, i.e., $a = qb \in (b)$.

Side remark: Note that in our terminology a "norm" $N : R \to \mathbb{N}$ is multiplicative, and indeed in our present application we are working with such norms. However, the multiplicativity property was not used in the proof. If R is a domain, let us define a **generalized Euclidean norm** on R to be a function $N : R \to \mathbb{N}$ such

that $N(r) = 0 \iff r = 0$ and such that for all $a \in R$, $b \in R \setminus \{0\}$, there exist $q, r \in R$ with $a = qb + r$ and $N(r) < N(b)$. Then what we have actually shown is that any domain which admits a generalized Euclidean norm is a PID.[5]

4.2. PIDs and UFDs.

One also knows that any PID is a UFD. This is true in general, but in the general case it is somewhat tricky to establish the existence of a factorization into irreducibles. In the presence of a multiplicative norm function $N : R \to \mathbb{N}$ – i.e., a function such that $N(x) = 0 \iff x = 0$, $N(x) = 1 \iff x \in R^\times$, $N(xy) = N(x)N(y) \forall x, y \in R$ – this part of the argument becomes much easier to establish, since for any nontrivial factorization $x = yz$ we have $N(y), N(z) < N(x)$. Complete details are available in *loc. cit.*

4.3. Some Euclidean quadratic rings.

Finally, we will show that our norm function on $\mathbb{Z}[\sqrt{-1}]$ is Euclidean. At this point it costs nothing extra, and indeed is rather enlightening, to consider the more general case of $\mathbb{Z}[\sqrt{D}]$ endowed with the norm function $N(a + b\sqrt{D}) = |a^2 - Db^2|$. According to the characterization of (multiplicative) Euclidean norms in the previous subsection, what we must show is: for all $\alpha \in \mathbb{Q}(\sqrt{D})$, there exists $\beta \in \mathbb{Z}[\sqrt{D}]$ with $N(\alpha - \beta) < 1$. A general element of α is of the form $r + s\sqrt{D}$ with $r, s \in \mathbb{Q}$, and we are trying to approximate it by an element $x + y\sqrt{D}$ with $x, y \in \mathbb{Z}$.

Let us try something easy: take x (resp. y) to be an integer nearest to r (resp. s). If z is any real number, there exists an integer n with $|z - n| \leq \frac{1}{2}$, and this bound is sharp, attained for all real numbers with fractional part $\frac{1}{2}$.[6] So let $x, y \in \mathbb{Z}$ be such that $|r - x|, |s - y| \leq \frac{1}{2}$. Is then $\beta = x + y\sqrt{D}$ a good enough approximation to $\alpha = r + s\sqrt{D}$? Consider the following quick and dirty estimate:

$$N(\alpha - \beta) = |(r - x)^2 - D(s - y)^2| \leq |r - x|^2 + |D||s - y|^2 \leq \frac{|D| + 1}{4}. \tag{20}$$

Evidently $\frac{|D|+1}{4} < 1$ iff $|D| < 3$.

So $D = -1$ works – i.e., the norm N on $\mathbb{Z}[\sqrt{-1}]$ is Euclidean, so $\mathbb{Z}[\sqrt{-1}]$ is a UFD, which fills in the gap in our proof of Theorem 3.4.

Also $D = 2$ and $D = -2$ work: the rings $\mathbb{Z}[\sqrt{-2}]$ and $\mathbb{Z}[\sqrt{2}]$ are UFDs.

It is natural to wonder whether the quick and dirty estimate can be improved. We have already seen that it *cannot* be for $D < -2$, since we already know that for such D, $\mathbb{Z}[\sqrt{D}]$ is not a UFD. This can be confirmed as follows: when $D < 0$, $N(a + b\sqrt{D}) = a^2 + Dy^2 = ||a + \sqrt{|D|}i||^2$; that is, our norm function is simply

[5]In fact further generalization is possible: in order for this simple argument to go through it is not necessary for the codomain of the norm function to be the natural numbers, but only a well-ordered set!

[6]Moreover when $|z - n| < \frac{1}{2}$, the nearest integer is unique, whereas for half-integral real numbers there are evidently two nearest integers. This does not matter to us: in this case take either nearest integer, i.e., either $z - \frac{1}{2}$ or $z + \frac{1}{2}$.

the square of the usual Euclidean length function evaluated on the complex number $a + \sqrt{|D|}i$. Moreover, in this case the ring $\mathbb{Z}[\sqrt{D}]$ lives naturally inside $\mathbb{C}$ as a *lattice*, whose fundamental parallelogram is simply a rectangle with sides 1 and $\sqrt{D}$. The problem now is to find, for a given point $z = a + bi \in \mathbb{C}$ with $a, b \in \mathbb{Q}$, the closest lattice point. But it is geometrically clear that the points which are furthest away from lattice points are precisely those which lie at the center of the corresponding rectangles, e.g. $\frac{1}{2} + \frac{1}{2}|D|i$. This shows that nothing was lost in our quick and dirty estimate.

The situation $D < 0$ is quite different, most of all because the geometric picture is different: $\mathbb{Z}[\sqrt{D}]$ now lives inside $\mathbb{R}$, but unlike in the previous case, it is not discrete. Rather, (Exercise X.X) it is *dense*: any interval (a, b) in $\mathbb{R}$ with $a < b$ contains some point $\alpha \in \mathbb{Z}[\sqrt{D}]$. So it is not clear that the above "coordinatewise nearest integer approximation" is the best possible approximation.

But even keeping the same approximating element β as above, we lose ground in the estimate $|(\frac{1}{2})^2 - D(\frac{1}{2})^2| = |\frac{1}{4} - \frac{D}{4}| \leq \frac{D+1}{4}$. Rather, we want $|\frac{1}{4} - \frac{D}{4}| < 1$, or $|D - 1| < 4$, so $D = 3$ also works. Thus, $\mathbb{Z}[\sqrt{3}]$ has a Euclidean norm, hence is a UFD. In summary, we get the following "bonus theorem":

THEOREM 3.8. *a) A prime p is of the form $x^2 + 2y^2$ iff -2 is a square modulo p.*
b) A prime p is of the form $|x^2 - 2y^2|$ iff 2 is a square modulo p.
c) A prime p is of the ofrm $|x^2 - 3y^2|$ iff 3 is a square modulo p.

Of course this brings attention to the fact that for an integer D, we do not know how to characterize the set of primes p such that D is a square mod p, except in the (easiest) case $D = -1$. The desire to answer this question is an excellent motivation for the **quadratic reciprocity law**, coming up shortly.

5. Composites of the Form $x^2 - Dy^2$

Now that we have determined which primes are of the form $x^2 + y^2$, it is natural to attempt to determine all nonzero integers which are sums of two squares.

An honest approach to this problem would begin by accumulating data and considering various special cases. Here we must unfortunately be somewhat more succinct.

Somewhat more generally, fix D a squarefree integer as before, and put

$$\mathcal{S}_D = \{n \in \mathbb{Z} \setminus \{0\} \mid \exists x, y \in \mathbb{Z},\ n = x^2 - Dy^2\},$$

the set of all nonzero integers of the form $x^2 - Dy^2$. Because of the multiplicativity of the norm function – or more precisely, the function $x + y\sqrt{D} \mapsto x^2 - Dy^2$, which takes on negative values when $D > 0$ – the subset $\mathcal{S}_D$ is closed under multiplication.

Remark: Certainly $1 \in \mathcal{S}_D$: $1 = 1^2 - D \cdot 0^2$. Therefore the multiplicative property can be rephrased by saying that $\mathcal{S}_D$ is a **submonoid** of the monoid $(\mathbb{Z} \setminus \{0\}, \cdot)$.

Now we know that the following positive integers are all sums of two squares: 1, 2, and prime $p \equiv 1 \pmod 4$, and n^2 for any integer n: $n^2 = (n)^2 + 0^2$. Now let

n be any positive integer, and write $p_1, \ldots, p_r$ for the distinct prime divisors of n which are congruent to 1 modulo 4, and $q_1, \ldots, q_s$ for the distinct prime divisors of n which are congruent to -1 modulo 4, so that

$$n = 2^a p_1^{m_1} \cdots p_r^{m_r} q_1^{n_1} \cdots q_s^{n_s},$$

for $a, m_1, \ldots, m_r, n_1, \ldots, n_s \in \mathbb{N}$. It follows that so long as $n_1, \ldots, n_s$ are all even, n is a product of sums of two squares and therefore itself a sum of two squares.

Finally, we wish to show that we have found all positive integers which are a sum of two squares.[7] Specifically, what we wish to show is that if $n \in \mathbb{Z}^+$ is a sum of two squares, then for any prime number $p \equiv -1 \pmod 4$, then $\mathrm{ord}_p(n)$ is even. For this it suffices to show the following

LEMMA 3.9. *Let $p \equiv -1 \pmod 4$ be a prime number, and suppose that there exist $x, y \in \mathbb{Z}$ such that $p \mid x^2 + y^2$, then $p \mid x$ and $p \mid y$.*

Before proving Lemma 3.9 let us show how it helps us. Indeed, suppose that a positive integer n is a sum of two squares: $n = x^2 + y^2$. Let p be any prime congruent to $-1 \pmod 4$, and assume that $p \mid n$ (otherwise $\mathrm{ord}_p(n) = 0$, which is even). Then by Lemma 3.9 $p \mid x$ and $p \mid y$, so that $\frac{n}{p^2} = (\frac{x}{p})^2 + (\frac{y}{p})^2$ expresses $\frac{n}{p^2}$ as a sum of two integral squares. But now we can repeat this process of repeated division by p^2 until we get an integer $\frac{n}{p^{2k}}$ which is not divisible by p. Thus $\mathrm{ord}_p(n) = 2k$ is even.

Proof of Lemma 3.9: It follows from the proof of the Two Squares Theorem that if $p \equiv -1 \pmod 4$ is a prime number, then it remains irreducible in $\mathbb{Z}[\sqrt{-1}]$. Let us recall why: otherwise $p = \alpha\beta$ with $N(\alpha), N(\beta) > 1$, and then taking norms gives

$$p^2 = N(p) = N(\alpha\beta) = N(\alpha)N(\beta),$$

and thus $N(\alpha) = N(\beta) = p$. Writing $\alpha = a + b\sqrt{-1}$, we get $p = N(\alpha) = a^2 + b^2$, so p is a sum of two squares, contrary to Fermat's Lemma.

Since $\mathbb{Z}[\sqrt{-1}]$ is a UFD, the irreducible element p is a prime element, hence Euclid's Lemma applies. We have $p \mid x^2 + y^2 = (x + y\sqrt{-1})(x - y\sqrt{-1})$, so that $p \mid x + y\sqrt{-1}$ or $p \mid x - y\sqrt{-1}$. This implies that $\frac{x}{p}, \frac{y}{p} \in \mathbb{Z}$, i.e., $p \mid x$ and $p \mid y$.

In summary, we have shown:

THEOREM 3.10. *(Full Two Squares Theorem) A positive integer n is a sum of two squares iff $\mathrm{ord}_p(n)$ is even for all primes $p \equiv -1 \pmod 4$.*

In that Lemma 3.9 uses (only) that $\mathbb{Z}[\sqrt{-1}]$ is a UFD, similar reasoning applies to other Euclidean quadratic rings $\mathbb{Z}[\sqrt{D}]$. In particular, there is a direct analogue of Theorem 3.10 for $x^2 + 2y^2$, which we will postpone until we determine for which primes p -2 is a square modulo p. When D is positive, the distinction between $\alpha\overline{\alpha} = a^2 - Db^2$ and $N(\alpha) = |a^2 - Db^2|$ becomes important: for instance as above we obviously have $1 \in \mathcal{S}_D$, but whether $-1 \in \mathcal{S}_D$ is a surprisingly difficult problem: to this day there is not a completely satisfactory solution.

In contrast, if $\mathbb{Z}[\sqrt{D}]$ is not a UFD, then even if we know which primes are of the form $x^2 - Dy^2$, we cannot use the above considerations to determine the set

[7] Why would we think this? Again, trial and error experimentation is the honest answer.

$\mathcal{S}_D$. For example take $D = -5$. Certainly $2 \neq x^2 + 5y^2$ and $3 \neq x^2 + 5y^2$, but $2 \cdot 3 = 1^2 + 5 \cdot 1^2$.

CHAPTER 4

Quadratic Reciprocity

We now come to the most important result in our course: the law of quadratic reciprocity, or, as Gauss called it, the **aureum theorema** ("golden theorem").

Many beginning students of number theory have a hard time appreciating this golden theorem. I find this quite understandable, as many first courses do not properly prepare for the result by discussing enough of the earlier work which makes quadratic reciprocity an inevitable discovery and its proof a cause for celebration. Happily, our study of quadratic rings and the quadratic form $x^2 - Dy^2$ has provided excellent motivation. There are also other motivations, involving (what we call here) the direct and inverse problems regarding the Legendre symbol.

A faithful historical description of the QR law is especially complicated and will not be attempted here; we confine ourselves to the following remarks. The first traces of QR can be found in Fermat's Lemma that -1 is a square modulo an odd prime p iff $p \equiv 1 \pmod 4$, so date back to the mid 1600's. Euler was the first to make conjectures equivalent to the QR law, in 1744. He was unable to prove most of his conjectures despite a steady effort over a period of about 40 years. Adrien-Marie Legendre was the first to make a serious attempt at a proof of the QR law, in the late 1700's. His proofs are incomplete but contain much valuable mathematics. He also introduced the Legendre symbol in 1798, which as we will see, is a magical piece of notation with advantages akin to Leibniz's dx in the study of differential calculus and its generalizations. Karl Friedrich Gauss gave the first complete proof of the QR law in 1797, at the age of 19(!). His argument used mathematical induction(!!). The proof appears in his groundbreaking work *Disquisitiones Arithmeticae* which was written in 1798 and first published in 1801.

The circle of ideas surrounding quadratic reciprocity is so rich that I have found it difficult to "linearize" it into one written presentation. (In any classroom presentation I have found it useful to begin each class on the subject with an inscription of the QR Law on a side board.) In the present notes, the ordering is as follows. In §1 we give a statement of the quadratic reciprocity law and its two supplements in elementary language. Then in §2 we discuss the Legendre symbol, restate QR in terms of it, and discuss (with proof) some algebraic properties of the Legendre symbol which are so important that they should be considered part of the quadratic reciprocity package. In §3 we return to our "unfinished theorems" about representation of primes by $|x^2 - Dy^2|$ when $\mathbb{Z}[\sqrt{D}]$ is a PID: using quadratic reciprocity, we can state and prove three **bonus theorems** which complement Fermat's Two Squares Theorem. In §4 we define and discuss the "direct and inverse problems" for the Legendre symbol and show how quadratic reciprocity is useful for both of these, in particular for rapid computation of Legendre symbols. More precisely, the computation would be rapid if we could somehow avoid having to

factor numbers quickly, and §5 explains how we can indeed avoid this by using an extension of the Legendre symbol due to Jacobi.

1. Statement of Quadratic Reciprocity

Notational comment: when we write something like $p \equiv a, b, c \pmod{n}$, what we mean is that $p \equiv a \pmod{n}$ or $p \equiv b \pmod{n}$ or $p \equiv c \pmod{n}$. (I don't see any other vaguely plausible interpretation, but it doesn't hurt to be careful.)

THEOREM 4.1. *(Quadratic Reciprocity Law) Let $p \neq q$ be odd primes. Then: (i) If $p \equiv 1 \pmod{4}$ or $q \equiv 1 \pmod{4}$, p is a square mod q iff q is a square mod p. (ii) If $p \equiv q \equiv 3 \pmod{4}$, p is a square mod q iff q **is not** a square mod p.*

THEOREM 4.2. *(First Supplement to the Quadratic Reciprocity Law) If p is an odd prime, then -1 is a square modulo p iff $p \equiv 1 \pmod{4}$.*

THEOREM 4.3. *(Second Supplement to the Quadratic Reciprocity Law) If p is an odd prime, then 2 is a square modulo p iff $p \equiv 1, 7 \pmod{8}$.*

2. The Legendre Symbol

2.1. Defining the Legendre Symbol.

We now introduce a piece of notation created by Adrien-Marie Legendre in 1798. There is no new idea here; it is "merely notation", but is an example of how incredibly useful well-chosen notation can be.

For n an integer and p an odd prime, we define the **Legendre symbol**

$$\left(\frac{n}{p}\right) := \begin{cases} 0, & \text{if } n \equiv 0 \pmod{p} \\ 1, & \text{if } n \bmod p \text{ is a nonzero square} \\ -1, & \text{if } n \bmod p \text{ is nonzero and not a square} \end{cases}$$

We must of course distinguish the Legendre symbol $\left(\frac{n}{p}\right)$ from the rational number $\frac{n}{p}$. To help with this, I recommend that $(\frac{n}{p})$ be read "n on p".[1]

Example 1: To compute $(\frac{12}{5})$, we must first observe that 5 does not divide 12 and then determine whether 12 is a nonzero square modulo 5. Since $12 \equiv 2 \pmod{5}$ and the squares modulo 5 are $1, 4$, the answer to the question "Is 12 a nonzero square modulo 5?" is negative, so $(\frac{12}{5}) = -1$.

Example 2: To compute $(\frac{101}{97})$ – note that 97 is prime! – we observe that 97 does not divide 101. Since $101 \equiv 4 \equiv 2^2 \pmod{97}$, the answer to the question "Is 101 a nonzero square modulo 97?" is positive, so $(\frac{101}{97}) = 1$.

Example 3: To compute $(\frac{97}{101})$ – note that 101 is prime! – we observe that 101 certainly does not divide 97. However, at the moment we do not have a very efficient way to determine whether 97 is a square modulo 101: our only method is to compute all of the squares modulo 101. Some calculation reveals that $400 = 20^2 = 3 \cdot 101 + 97$, so $20^2 \equiv 97 \pmod{101}$. Thus 97 is indeed a square modulo 101, so $(\frac{97}{101}) = 1$.

[1] There is in fact *some* relationship with "n divided by p": if we divide n by p, getting $n = qp + r$ with $0 \leq r < p$, then the Legendre symbols $(\frac{n}{p})$ and $(\frac{r}{p})$ are equal.

2.2. Restatement of Quadratic Reciprocity.

THEOREM 4.4. *(Quadratic Reciprocity) Let p and q be distinct odd primes.*

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

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

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

2.3. Some elementary group theory related to the Legendre symbol.

Let p be an odd prime, and consider the group $U(p) = (\mathbb{Z}/p\mathbb{Z})^\times$; since p is prime, this is precisely the multiplicative group of nonzero elements of $\mathbb{Z}/p\mathbb{Z}$ under multiplication: in particular, it is a finite commutative group of even order $p-1$.

In fact $U(p)$ is a cyclic group: there exists some element $g \in U(p)$ such that every element $x \in U(p)$ is of the form g^i for a unique $0 \leq i < p$. In classical number-theoretic language the element g (often viewed as an integer, $0 < g < p$) is a **primitive root modulo p**. This is nontrivial to prove. We do in fact give the proof elsewhere in these notes, but in at least one version of the course, we are covering quadratic reciprocity before the material on the Euler φ function which we use in our proof of this fact. So we would like to give a more elementary discussion of some weaker properties of $U(p)$ that suffice for our needs here.

Let $(G, \cdot)$ be a commutative group, and let n be a positive integer. The map

$$[n] : G \to G,\ x \mapsto x^n$$

which sends each element to its nth power, is a homomorphism. We denote the kernel of the map by $G[n]$; this is the subgroup of all elements of order dividing n, often called the **n-torsion subgroup** of G. We put $G^n = [n](G)$, the image of the homomorphism, which is the subgroup of elements of G which are nth powers. There is thus a canonical isomorphism

$$[n] : G/G[n] \xrightarrow{\sim} G^n.$$

Now further suppose that G is finite. Then

$$\#G^n = \frac{\#G}{\#G[n]}.$$

Consider for a moment the case $\gcd(n, \#G) = 1$. Suppose $g \in G[n]$. Then the order of g divides n, whereas by Lagrange's theorem, the order of g divides $\#G$, so the order of G divides $\gcd(n, \#G) = 1$: so $g = 1$ and $G[n] = \{1\}$. Thus $\#G^n = \#G$ so $G^n = G$. So in this case every element of G is an nth power.

We remark in passing that the converse is also true: if $\gcd(n, \#G) > 1$, then $G[n]$ is nontrivial, so the subgroup G^n of nth powers is proper in G. We do not need this general result, so we do not prove it here, but mention only that it can be deduce from the classification theorem for finite commutative groups.

Now we specialize to the case $G = U(p) = (\mathbb{Z}/p\mathbb{Z})^\times$ and $n = 2$. Then

$$G[2] = \{x \in \mathbb{Z}/p\mathbb{Z} \setminus \{0\} \mid x^2 = 1\}.$$

We claim that $G[2] = \{\pm 1\}$. First, note that since p is odd, $1 \not\equiv -1 \pmod{p}$, i.e., $+1$ and -1 are distinct elements in $\mathbb{Z}/p\mathbb{Z}$, and they clearly both square to 1, so that $G[2]$ contains at least the two element subgroup $\{\pm 1\}$. Conversely, as above every element of $G[2]$ is a root of the quadratic polynomial $t^2 - 1$ in the field $\mathbb{Z}/p\mathbb{Z}$. But a polynomial of degree d over any field (or integral domain) can have at most d distinct roots: whenever $p(a) = 0$, applying the division algorithm to $p(t)$ and $t - a$ gives $p(t) = q(t)(t-a) + c$, where c is a constant, and plugging in $t = a$ gives $c = 0$. Thus we can factor out $t - a$ and the degree decreases by 1. Therefore $\#G[2] \leq 2$, and since we have already found two elements, we must have $G[2] = \{\pm 1\}$.

So G^2 is an index two subgroup of G and the quotient G/G^2 has order two. Like any group of order 2, it is uniquely isomorphic to the group $\{\pm 1\}$ under multiplication. Thus we have defined a surjective group homomorphism

$$L : U(p) \to \{\pm 1\},$$

namely we take $x \in U(p)$ to the coset $xU(p)^2$. So, $L(x) = 1$ if x is a square in $(\mathbb{Z}/p\mathbb{Z})^\times$ and $L(x) = -1$ otherwise. But this means that for all $x \in \mathbb{Z}/p\mathbb{Z} \setminus \{0\}$, $L(x) = (\frac{x}{p})$. Thus we have recovered the Legendre symbol in terms of purely algebraic considerations and also shown that

$$\forall x, y \in U(p), \left(\frac{xy}{p}\right) = \left(\frac{x}{p}\right)\left(\frac{y}{p}\right).$$

In fact we can give a (useful!) second description of the Legendre symbol using power maps. (This discussion repeats the proof of Proposition 3.2, but we are happy to do so.) To see this, consider the map

$$[\frac{p-1}{2}] : U(p) \to U(p).$$

We claim that the kernel of this map is again the subgroup $U(p)^2$ of squares, of order $\frac{p-1}{2}$. On the one hand, observe that $U(p)^2 \subset U(p)[\frac{p-1}{2}]$: indeed $(x^2)^{\frac{p-1}{2}} = x^{p-1} = 1$ by Lagrange's Theorem. Conversely, the elements of $U(p)[\frac{p-1}{2}]$ are roots of the polynomial $t^{\frac{p-1}{2}} - 1$ in the field $\mathbb{Z}/p\mathbb{Z}$, so there are at most $\frac{p-1}{2}$ of them. Thus $U(p)^2 = U(p)[\frac{p-1}{2}]$. By similar reasoning we have $U(p)^{\frac{p-1}{2}} \subset \{\pm 1\}$, hence we can view $[\frac{p-1}{2}]$ as a homomorphism

$$L' = [\frac{p-1}{2}] : U(p) \to \{\pm 1\}.$$

Since the kernel of L' is precisely the subgroup $U(p)^2$ and there are only two possible values, it must be the case that $L'(x) = -1$ for all $x \in U(p) \setminus U(p)^2$. In other words, we have $L'(x) = (\frac{x}{p})$.

The following result is essentially a summary of the above work. We strongly recommend that the reader take time out to convince herself of this.

PROPOSITION 4.5. *The following hold for any $a, b \in \mathbb{Z}$ and any odd prime p.*

a) $\left(\frac{a}{p}\right)$ *depends only on the residue class of a modulo p.*

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

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

Note that by taking $a = -1$ in Proposition 4.5b), we get

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

This is precisely the First Supplement to the quadratic reciprocity law, which we have now proved twice (in the handout on Pythagorean triples we called it **Fermat's Lemma** and proved it using Wilson's theorem).

2.4. A faster proof using the cyclicity of $U(p)$.

If we happen to know that the unit group $U(p) = (\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic, we can give a much more streamlined proof of Proposition 4.5. First note that part a) is obvious from the definition. Moreover, if we assume part b), part c) follows immediately:

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

So it remains to prove part b). But now suppose that g is a generator for the group $U(p)$, so that we can write $a = g^i$. Then $a^{\frac{p-1}{2}} = g^{\frac{i(p-1)}{2}}$.
Case i: i is even. Then on the one hand $a = (g^{\frac{i}{2}})^2$ is a square in $U(p)$. On the other hand $p - 1 \mid i\frac{p-1}{2}$, so that $g^{\frac{i(p-1)}{2}} = 1$ by Lagrange's theorem.
Case 2: i is odd. Then on the on the one hand $a = g^i$ is not a square in $U(p)$: for instance, we know that the subgroup of squares has exactly $\frac{p-1}{2}$ elements, and we found $\frac{p-1}{2}$ distinct elements in Case 1 above: $\{g^{2k} \mid 0 \leq k < \frac{p-1}{2}\}$. On the other hand, since i is odd, $p - 1 \nmid i\frac{p-1}{2}$, so that $a^{\frac{p-1}{2}} = g^{\frac{i(p-1)}{2}} \neq 1$. Since its square is 1, it must therefore be equal to -1.

3. Motivating Quadratic Reciprocity I: Bonus Theorems

3.1. Some unfinished theorems.

An excellent motivation for the quadratic reciprocity law is provided by our previous study of the equation $x^2 - Dy^2 = p$. Recall we have proved:

THEOREM 4.6. *Let D be squarefree integer different from 0 and 1. Assume that the ring $\mathbb{Z}[\sqrt{D}]$ is a UFD. Then, for a prime number p, TFAE:*
(i) There exist $x, y \in \mathbb{Z}$ such that $p = |x^2 - Dy^2|$.
(ii) There exists $x \in \mathbb{Z}$ such that $D \equiv x^2 \pmod{p}$.

Moreover we know that $\mathbb{Z}[\sqrt{D}]$ is a UFD when $D \in \{-1, \pm 2, 3\}$. The case $D = -1$ yielded Fermat's two squares theorem given the additional knowledge that -1 is a square modulo an odd prime p iff $p \equiv 1 \pmod 4$. To complete our "bonus theorems" we need answers to the following questions:

- For which odd primes p is it the case that -2 is a square modulo p?
- For which odd primes p is it the case that 2 is a square modulo p?
- For which odd primes p is it the case that 3 is a square modulo p?

Comparing with the answer for $D = -1$, one might hope that the answer is in terms of some congruence condition on p. Let's look at some data:

The odd primes $p < 200$ for which -2 is a square modulo p are:

$$3, 11, 17, 19, 41, 43, 59, 67, 73, 83, 89, 97, 107, 113, 131, 137, 139, 163, 179, 193.$$

Notice that these are precisely the primes $p < 200$ with $p \equiv 1, 3 \pmod 8$.

For $D = 2, 3$ we will give some data and allow you a chance to find the pattern.

The odd primes $p < 200$ for which 2 is a square modulo p are:

$$7, 17, 23, 31, 41, 47, 71, 73, 79, 89, 97, 103, 113, 127, 137, 151, 167, 191, 193, 199.$$

The odd primes $p < 200$ for which 3 is a square modulo p are:

$$3, 11, 13, 23, 37, 47, 59, 61, 71, 73, 83, 97, 107, 109, 131, 157, 167, 179, 181, 191, 193.$$

While we are at it, why not a bit more data?

The odd primes $p < 200$ for which 5 is a square modulo p are:

$$5, 11, 19, 29, 31, 41, 59, 61, 71, 79, 89, 101, 109, 131, 139, 149, 151, 179, 181, 191, 199.$$

The odd primes $p < 200$ for which 7 is a square modulo p are:

$$3, 7, 19, 29, 31, 37, 47, 53, 59, 83, 103, 109, 113, 131, 137, 139, 149, 167, 193, 197, 199.$$

3.2. With the help of quadratic reciprocity.

We already know that a prime p is of the form $|x^2 - 2y^2|$ iff $(\frac{2}{p}) = 1$, and the second supplement tells us that this latter conditions holds iff $p \equiv 1, 7 \pmod 8$. While we are here, let's deal with the absolute value: it happens that $\mathbb{Z}[\sqrt{2}]$ contains an element of norm -1, namely $1 - \sqrt{2}$:

$$N(1 - \sqrt{2}) = (1 - \sqrt{2})(1 + \sqrt{2}) = 1^2 - 2 \cdot 1^2 = -1.$$

From this and the multiplicaitivity of the norm map, it follows that if we can represent any integer n in the form $x^2 - 2y^2$, we can also represent it in the form $-(x^2 - 2y^2)$, and conversely. From this it follows that the absolute value is superfluous and we get the following result.

THEOREM 4.7. *(First Bonus Theorem) A prime number p is of the form $x^2 - 2y^2$ iff $p = 2$ of $p \equiv 1, 7 \pmod 8$.*

Now let's look at the case of $D = -2$, i.e., the form $x^2 + 2y^2$. Since $2 = 0^2 + 2 \cdot 1^2$, 2 is of of the form $x^2 + 2y^2$. Now assume that p is odd. We know that an odd prime p is of the form $x^2 + 2y^2$ iff $(\frac{-2}{p}) = 1$. We don't have a single law for this, but the multiplicativity of the Legendre symbol comes to our rescue. Indeed,

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

so

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

Case 1: $(\frac{-1}{p}) = (\frac{2}{p}) = 1$. By the first and second supplements, this occurs iff $p \equiv 1 \pmod 4$ and $p \equiv 1, 7 \pmod 8$, so iff $p \equiv 1 \pmod 8$.

Case 2: $(\frac{-1}{p}) = (\frac{2}{p}) = -1$. By the first and second supplements, this occurs iff $p \equiv 3 \pmod 4$ and $p \equiv 3, 5 \pmod 8$, so iff $p \equiv 3 \pmod 8$. Thus:

THEOREM 4.8. *(Second Bonus Theorem) A prime number p is of the form $x^2 + 2y^2$ iff $p = 2$ or $p \equiv 1, 3 \pmod 8$.*

Now let's look at the case of $D = 3$, i.e., the form $|x^2 - 3y^2|$. We know that a prime p is of this form iff $(\frac{3}{p}) = 1$. Now we use QR itself, and there are two cases:
Case 1: If $p \equiv 1 \pmod 4$, then $(\frac{3}{p}) = 1$ iff $p \equiv 1 \pmod 3$.
Case 2: If $p \equiv 3 \pmod 4$, then $(\frac{3}{p}) = 1$ iff $p \equiv -1 \pmod 3$.

The congruence conditions can be consolidated by going mod 12. We get that $(\frac{3}{p}) = 1$ iff $p \equiv 1, 11 \pmod{12}$. Again we can ask what happens when we try to remove the absolute value. This time things work out somewhat differently.

THEOREM 4.9. *(Third Bonus Theorem) For a prime p, the equation $x^2 - 3y^2 = p$ has an integral solution iff $p \equiv 1 \pmod{12}$. The equation $3y^2 - x^2 = p$ has an integral solution iff $p = 2$, $p = 3$ or $p \equiv 11 \pmod{12}$.*

PROOF. First we deal with the two exceptional cases. Suppose $p = 2$: reducing $x^2 - 3y^2 = 2$ modulo 3, we get $x^2 \equiv 2 \pmod 3$, which we know has no solution. Note that on the other hand $3(1)^2 - 1^2 = 2$, so 2 is of the form $3y^2 - x^2$. Now suppose $p = 3$: reducing $x^2 - 3y^2 = 3$ modulo 4, we get $x^2 - 3y^2 \equiv x^2 + y^2 \equiv 3 \pmod 4$, which (as we have seen before) has no integral solution. On the other hand, $3 = 3(1)^2 - 0^2$, so 3 is of the form $3y^2 - x^2$.

Now suppose that $p > 3$. Since $\mathbb{Z}[\sqrt{3}]$ is a PID, we know that p is of the form $p = |x^2 - 3y^2|$ iff 3 is a square modulo p, i.e., iff $p = 3$ or $(\frac{3}{p}) = 1$. By quadratic reciprocity, this last condition can be expressed as a congruence modulo $4 \cdot 3 = 12$, specifically $p \equiv \pm 1 \pmod{12}$. So if $p \equiv 1, 11 \pmod{12}$ then at least one of the following holds:

$$p = x^2 - 3y^2 \tag{21}$$

or

$$p = 3y^2 - x^2. \tag{22}$$

It turns out that for any prime p, exactly one of the two equations (21), (22) holds, which is extremely convenient: it means that we can always show that one of the equations holds by showing that the other one does not hold!

Indeed, if we reduce the equation $p = x^2 - 3y^2$ modulo 3: we get $p \equiv x^2 \pmod 3$, i.e., $(\frac{p}{3}) = 1$, so $p \equiv 1 \pmod 3$. So if $p \equiv 11 \pmod{12}$ then p is not of the form $x^2 - 3y^2$ so must be of the form $3y^2 - x^2$. Simiarly, if we reduce the equation $p = 3y^2 - x^2$ modulo 3, we get $p \equiv -x^2 \equiv -1 \pmod 3$, so if $p \equiv 1 \pmod 3$ then 22 has no solution, so it must be that $p = x^2 - 3y^2$ does have a solution. □

A very similar argument establishes the following more general result.

THEOREM 4.10. *Suppose $q \equiv 3 \pmod 4$ is a prime such that $\mathbb{Z}[\sqrt{q}]$ is a PID. Then the equation $x^2 - qy^2 = p$ has a solution iff $p \equiv 1 \pmod 4$ and $(\frac{p}{q}) = 1$.*

3.3. Auxiliary congruences.

The restriction to $q \equiv 3 \pmod 4$ in Theorem 4.10 may appear artificial. But those who have done their homework know better: in fact if $\mathbb{Z}[\sqrt{q}]$ is a PID, then we must have $q = 2$ (which we have already discussed) or $q \equiv 3 \pmod 4$. (Otherwise $\mathbb{Z}[\sqrt{q}]$ is not integrally closed.) A closer look reveals that the distinction between primes which are 1 (mod 4) and primes which are 3 (mod 4) is a central, albeit somewhat mysterious part, of the natural behavior of quadratic forms.

One way to see this is in terms of what I shall call **auxiliary congruences**. Namely, in our initial study of the equation $|x^2 - Dy^2| = p$, we did not consider all possible congruence obstructions (as e.g. in Legendre's Theorem) but only condition that we got upon reducing modulo p: namely that D is a square modulo p. Notice that we could also reduce modulo D to get some further conditions: more on this in a moment. But why didn't we reduce modulo D before? The simple but strange answer is that we simply didn't need to: it happened that when $\mathbb{Z}[\sqrt{D}]$ is a PID, we were able to prove that the necessary condition that D be a square modulo p was also sufficient for p to be of the form $|x^2 - Dy^2| = p$.

But this is rather surprising. Let's look closer, and to fix ideas let us take p and q distinct odd primes, and look at the equation

$$x^2 + qy^2 = p.$$

Then reducing modulo p gives $(\frac{-q}{p}) = 1$, whereas reducing modulo q gives $(\frac{p}{q}) = 1$. How do these two conditions interact with each other? Let's examine the cases:

Case 1: $p \equiv 1 \pmod 4$. Then $(\frac{-q}{p}) = (\frac{-1}{p})(\frac{q}{p}) = (\frac{q}{p}) = (\frac{p}{q})$. So the conditions are **redundant**.

Example 4.11. *Take $q = 5$. Then the congruence conditions tell us that if $p \equiv 1 \pmod 4$ is of the form $x^2 + 5y^2$, we must have $(\frac{p}{5}) = 1$, i.e., $p \equiv 1, 4 \pmod 5$. Thus, every prime $p \equiv 1 \pmod 4$ which is represented by $x^2 + 5y^2$ lies in one of the two congruence classes $p \equiv 1, 9 \pmod{20}$. As we know, $\mathbb{Z}[\sqrt{-5}]$ is* not *a PID, so nothing we have proved tells us anything about the converse, but the above computations show that for all $p < 200$ we have: $p \equiv 1, 9 \pmod{20} \implies p = x^2 + 5y^2$. It is easy to extend the computations to check this for all primes up to say 10^6. In fact it is true, although the proof requires techniques that we have not developed.*

Example 4.12. *Take $q = 3$. Then the congruence conditions tell us that if $p \equiv 1 \pmod 4$ is of the form $x^2 + 3y^2$, then $p \equiv 1 \pmod 3$. Again computations support that every prime $p \equiv 1 \pmod 1 2$ is of the form $x^2 + 3y^2$.*

Case 2: $p \equiv 3 \pmod 4$. Then $(\frac{-q}{p}) = (\frac{-1}{p})(\frac{q}{p}) = -(\frac{q}{p})$. To compare this to the condition $(\frac{p}{q}) = 1$, we need to consider further cases.

Case 2a) Suppose also $q \equiv 1 \pmod 4$. Then $1 = (\frac{-q}{p}) = -(\frac{q}{p}) = -(\frac{p}{q})$, i.e., $(\frac{p}{q}) = -1$. This is **inconsistent** with $(\frac{p}{q}) = 1$, so we deduce that when $q \equiv 1 \pmod 4$, $p = x^2 + qy^2 \implies p \equiv 1 \pmod 4$.

This is a new phenomenon for us. Note that when $q = 5$, in conjunction with the above (unproved) result, we get the following

THEOREM 4.13. *An odd prime p is of the form $x^2 + 5y^2$ iff $p \equiv 1, 9 \pmod{20}$.*

Case 2b): Suppose also $q \equiv 3 \pmod 4$. Then $1 = (\frac{-q}{p}) = -(\frac{q}{p}) = (\frac{p}{q})$. Thus the two congruence conditions are **consistent** in this case.

EXAMPLE 4.14. *Let's reconsider $q = 3$. Nothing in our analysis ruled out a prime $p \equiv 3 \pmod 4$ (except $p = 3$) being of the form $x^2 + 3y^2$: the only congruence condition we found is the main one $1 = (\frac{-3}{p}) = \frac{p}{3}$, i.e., $p \equiv 1 \pmod 3$. In this case computations suggest that an odd prime p is of the form $x^2 + 3y^2$ iff $p \equiv 1 \pmod 3$. Note that this is exactly the result that we would have gotten if $\mathbb{Z}[\sqrt{3}]$ were a UFD except that then 2 would also be of the form $x^2 + 3y^2$, which was exactly what we used to see that $\mathbb{Z}[\sqrt{3}]$ isn't a UFD! It turns out that we* can *prove this result with the techniques we have: an argument is sketched in the exercises.*

These considerations have turned up more questions than answers. Our point is that the distinction between primes $p \equiv 1 \pmod 4$ and $p \equiv 3 \pmod 4$ is something that is embedded quite deeply into the behavior of quadratic rings and quadratic equations. A proper understanding of this phenomenon goes under the heading **genus theory**, which was treated by Gauss in his *Disquisitiones Arithmeticae* and is intimately related to contemporary issues in number theory.

4. Motivating Quadratic Reciprocity II: Direct and Inverse Problems

4.1. The direct and inverse problems.

We wish to discuss "reciprocal" problems concerning quadratic residues, which can be phrased in terms of whether we regard the Legendre symbol $\left(\frac{n}{p}\right)$ as a function of its numerator or as a function of its denominator.

DIRECT PROBLEMS: Fix an odd prime p.
DIRECT PROBLEM A: Determine all integers which are squares modulo p.
DIRECT PROBLM B: Determine whether a given integer n is a square modulo p.

By Proposition 4.5a), the answer only depends upon n modulo p, so for fixed p it is a finite problem: we know that exactly half of the elements of $\mathbb{F}_p^\times$ are squares, so for instance to compute all of them we could simply calculate $1^2, 2^2, \ldots, (p-1)^2$ modulo p.[2] However if p is large this will take a long time, and it is natural to wonder whether there is a faster way of computing $\left(\frac{n}{p}\right)$ for some specific n.

INVERSE PROBLEM: Fix $n \in \mathbb{Z}$. For which odd primes p is $\left(\frac{n}{p}\right) = 1$?

Example: The case $n = -1$ was needed to prove the two squares theorem. We found that $(\frac{-1}{p}) = 1$ iff $p \equiv 1 \pmod 4$. Note that, although our original proof was more elementary, this follows immediately from Proposition 4.5b): $(\frac{-1}{p}) = (-1)^{\frac{p-1}{2}}$.

In contrast to the direct problems, the inverse problem is apparently an *infinite*

[2]In fact this gives every square twice; we will get every square once by computing the squares up to $(\frac{p-1}{2})^2$, as we saw in the Handout on Sums of Two Squares.

problem. Moreover, the inverse problem comes up naturally in applications: indeed solving the inverse problem for $n = \pm 2, 3$ was exactly what we did in the last section in order to complete our study of the forms $x^2 - ny^2$.

4.2. With the help of quadratic reciprocity.

We now make two key observations. First: The Quadratic Reciprocity Law allows us to reduce the inverse problem to the direct problem A.

Example: Take $n = 5$. For which odd primes p is 5 a square modulo p?

Answer: Since 5 is 1 (mod 4), $\left(\frac{5}{p}\right) = \left(\frac{p}{5}\right)$, and we know what the squares are mod 5: ± 1. Thus the answer is that 5 is a square modulo p iff $p \equiv \pm 1 \pmod 5$.

Example: Take $n = 7$. For which odd primes p is 7 a square modulo p?

Answer: Since 7 is 3 (mod 4), we need to distinguish two cases: $p \equiv 1 \pmod 4$ and $p \equiv -1 \pmod 4$. If $p \equiv 1 \pmod 4$, then $\left(\frac{7}{p}\right) = \left(\frac{p}{7}\right)$, so we just want p to be a square modulo 7. The squares mod 7 are $1^2 \equiv 1, 2^2 \equiv 4$ and $3^2 \equiv 2$. We now have both a congruence condition mod 7 and a congruence condition mod 4: by the Chinese Remainder theorem, these conditions can be expressed by congruence conditions mod 28: namely we want $p \equiv 1, 9, 25 \pmod{28}$.

Next we consider the case $p \equiv -1 \pmod 4$. This time since p and 7 are both -1 mod 4, QR tells us that $\left(\frac{7}{p}\right) = -1\left(\frac{p}{7}\right)$, so we want the nonsquares modulo 7, or $3, 5, 6$. Again we may combine these with the congruence $p \equiv -1 \pmod 4$ by going mod 28, to get $p \equiv 3, 19, 27$. So 7 is a square modulo p iff

$$p \equiv 1,\ 3,\ 9,\ 19,\ 25,\ \text{or } 27 \pmod{28}.$$

The QR law leads to the following general solution of the inverse problem:

Corollary 4.15. *Let q be any odd prime.*
a) If $q \equiv 1 \pmod 4$, then $\left(\frac{q}{p}\right) = 1$ iff p is congruent to a square modulo q (so lies in one of $\frac{q-1}{2}$ residue classes modulo q).
b) If $q \equiv -1 \pmod 4$, then $\left(\frac{q}{p}\right) = 1$ iff $p \equiv \pm x^2 \pmod{4q}$ (so lies in one of $q-1$ out of the $2(q-1)$ reduced residue classes modulo $4q$).

Corollary 4.15 was first conjectured by Euler and is in fact equivalent to the QR law. As we will not be using Corollary 4.15 in the sequel, we leave the proof as an exercise.

So much for the first observation. Here is the second:

The QR law yields an efficient algorithm for Direct Problem B.

This is best explained by way of examples.

Example: Suppose we want to compute $\left(\frac{7}{19}\right)$. Using QR we can "invert" the

Legendre symbol, tacking on an extra factor of -1 because $7 \equiv 19 \equiv -1 \pmod 4$:

$$\left(\frac{7}{19}\right) = -\left(\frac{19}{7}\right) = -\left(\frac{5}{7}\right) = -\left(\frac{7}{5}\right) = -\left(\frac{2}{5}\right).$$

We have reduced to a problem we know: 2 is not a square mod 5, so the final answer is $\left(\frac{7}{19}\right) = -(-1) = 1$.

Example:

$$\left(\frac{41}{103}\right) = \left(\frac{103}{41}\right) = \left(\frac{21}{41}\right) = \left(\frac{3}{41}\right)\left(\frac{7}{41}\right) = \left(\frac{41}{3}\right)\left(\frac{41}{7}\right)$$
$$= \left(\frac{-1}{3}\right)\left(\frac{-1}{7}\right) = -1 \cdot -1 = 1.$$

Example:

$$\left(\frac{79}{101}\right) = \left(\frac{101}{79}\right) = \left(\frac{22}{79}\right) = \left(\frac{2}{79}\right)\left(\frac{11}{79}\right) =$$
$$1 \cdot \left(\frac{11}{79}\right) = -\left(\frac{79}{11}\right) = -\left(\frac{2}{11}\right) = -(-1) = 1.$$

Let us now stop and make an important observation: the quadratic reciprocity law along with its first and second supplements, together with parts a) and c) of Proposition 4.5, allows for a computation of the Legendre symbol $\left(\frac{n}{p}\right)$ in all cases. Indeed, it is multiplicative in the numerator, so we may factor n as follows:

$$n = (-1)^{\epsilon} 2^a p^b p_1 \cdots p_r \cdot m^2,$$

where $\epsilon = \pm 1$, the primes $p_1, \ldots, p_r$ are distinct and prime to p, and m is prime to p. If $b > 0$ then the symbol evaluates to 0. Otherwise we have

$$\left(\frac{n}{p}\right) = \left(\frac{-1}{p}\right)^{\epsilon} \left(\frac{2}{p}\right)^{a} \prod_i \left(\frac{p_i}{p}\right).$$

5. The Jacobi Symbol

Computing Legendre symbols via the method of the previous section is, for moderately small values of n and p, much faster and more pleasant to do by hand than computing the list of all $\frac{p-1}{2}$ quadratic residues mod p. However, when the numbers get larger a "hidden cost" of the previous calculation becomes important: the calculation requires us to do several factorizations, and factorization is the *ne plus ultra* of time-consuming number-theoretic calculations.

In fact it is not necessary to do any factorization at all, except to factor out the largest power of 2, which is computationally trivial (especially if the number is stored in binary form!). One can use a generalization of the Legendre symbol introduced in 1837 by Carl Gustav Jacob Jacobi (1804-1851).

For a an integer and b an odd positive integer, we define the **Jacobi symbol**

$$\left(\frac{a}{b}\right) = \left(\frac{a}{p_1}\right) \cdots \left(\frac{a}{p_r}\right),$$

where $b = p_1 \cdots p_r$ is the factorization of b into (not necessarily distinct!) primes.

Warning: If a is a square modulo b, then $\left(\frac{a}{b}\right) = 1$, but the converse does not hold (you are asked to supply a counterexample in the homework). The Jacobi symbol is instead a "formal" generalization of the Legendre symbol, as is summarized by the following two results:

PROPOSITION 4.16. *Let a, a_1, a_2 be integers and b, b_1, b_2 be odd positive integers.*
a) $\left(\frac{a_1}{b}\right) = \left(\frac{a_2}{b}\right)$ *if* $a_1 \equiv a_2 \pmod{b}$.
b) $\left(\frac{a_1 a_2}{b}\right) = \left(\frac{a_1}{b}\right)\left(\frac{a_2}{b}\right)$.
c) $\left(\frac{a}{b_1 b_2}\right) = \left(\frac{a}{b_1}\right)\left(\frac{a}{b_2}\right)$.

THEOREM 4.17. *(QR Law for the Jacobi Symbol) Let a be an integer and b an odd positive integer.*
a) $\left(\frac{-1}{b}\right) = (-1)^{\frac{b-1}{2}}$.
b) $\left(\frac{2}{b}\right) = (-1)^{\frac{b^2-1}{8}}$.
c) If a is also odd and positive then

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

The point is that the Jacobi symbol equals the Legendre symbol when the denominator is a prime, and is moreover completely determined by Proposition 4.16a) and Theorem 4.17. Therefore one can compute Legendre symbols by a process of repeated inversion and reduction of the numerator modulo the denominator, without worrying about whether the numerator or denominator is prime.

If a and b each have no more than k digits, then computing the Jacobi symbol $\left(\frac{a}{b}\right)$ using the QR law requires no more than a constant times k^2 steps, or more succinctly, can be done in time $O(k^2)$.[3] In particular, when $b = p$ is prime, the algorithm takes $O(\log^2 p)$ steps so is **polynomial time** (in the number of digits of p), whereas computing all $\frac{p-1}{2}$ quadratic residues takes time $O(p)$.

Using the Euler relation $(\frac{a}{p}) \equiv a^{\frac{p-1}{2}} \pmod{p}$ to compute $(\frac{a}{p})$ is also rather efficient, as one can takie advantage of a **powering algorithm** to rapidly compute exponents modulo p (the basic idea being simply to not compute the integer $a^{\frac{p-1}{2}}$ at all but rather to alternate raising a to successively larger powers and reducing the result modulo p): this can be done in time $O(\log^3 p)$. For more information on this and many other topics related to number-theoretic algorithms, we recommend Henri Cohen's *A Course in Computational Algebraic Number Theory*.

6. Preliminaries on Congruences in Cyclotomic Rings

For a positive integer n, let $\zeta_n = e^{\frac{2\pi i}{n}}$ be a primitive nth root of unity, and let

$$R_n = \mathbb{Z}[\zeta_n] = \{a_0 + a_1\zeta_n + \ldots + a_{n-1}\zeta_n^{n-1} \mid a_i \in \mathbb{Z}\}.$$

[3]The notation $O(f(x))$ is used in algorithmic complexity theory and also in analytic number theory to indicate a quantity which is bounded above by a constant times $f(x)$.

Recall that an **algebraic integer** is a complex number α which satisfies a monic polynomial relation with $\mathbb{Z}$-coefficients: there exist n and $a_0, \ldots, a_{n-1}$ such that

$$\alpha^n + a_{n-1}\alpha^{n-1} + \ldots + a_1\alpha + a_0.$$

We need the following purely algebraic fact:

PROPOSITION 4.18. *a) Every element of R_n is an algebraic integer.*
b) $R_n \cap \mathbb{Q} = \mathbb{Z}$.

PROOF. a) [**F**, Prop. 7.2] Let $z \in R_n$. For $0 \leq i \leq n-1$, $z\zeta_n^i$ is an element of R_n hence may be written as

$$z\zeta_n^i = \sum_{j=0}^{n-1} a_{ij}\zeta_n^j$$

for some $a_{ij} \in \mathbb{Z}$. Altogether we get an $n \times n$ matrix $A = (a_{ij})_{0 \leq i,j \leq n-1}$ with integer coefficients. Let $P(t) = \det(tI - A)$ be the *characteristic polynomial* of A. Since A has integer entries, $P(t)$ is monic with $\mathbb{Z}$-coefficients. Now, if we put $v = (1, \zeta_n, \zeta_n^2, \ldots, \zeta_n^{n-1})^T$ – i.e., we are viewing v as an $n \times 1$ matrix, or "column vector" – then the above equations are equivalent to the one matrix equation

$$Av = zv.$$

Thus the complex number z is an eigenvalue of the integer matrix A. Hence, by basic linear algebra, $P(z) = 0$, so z is an algebraic integer.
b) By Theorem 1.26, an algebraic integer which is also a rational number is an ordinary integer. □

Let p be a prime number; for $x, y \in R_n$, we will write $x \equiv y \pmod{p}$ to mean that there exists a $z \in R_n$ such that $x - y = pz$. Otherwise put, this is congruence modulo the principal ideal pR_n of R_n.

Since $\mathbb{Z} \subset R_n$, if x and y are ordinary integers, the notation $x \equiv y \pmod{p}$ is ambiguous: interpreting it as a usual congruence in the integers, it means that there exists an integer n such that $x - y = pn$; and interpreting it as a congruence in R_n, it means that $x - y = pz$ for some $z \in R_n$. The key technical point is that these two notions of congruence are in fact the same:

COROLLARY 4.19. *If $x, y \in \mathbb{Z}$ and $z \in R_n$ are such that $x - y = pz$, then $z \in \mathbb{Z}$.*

PROOF. Indeed $z = \frac{x-y}{p} \in R_n \cap \mathbb{Q} = \mathbb{Z}$. □

To prove the second supplement we will take $n = 8$. To prove the QR law we will take $n = p$ an odd prime. These choices will be constant throughout each of the proofs so we will abbreviate $\zeta = \zeta_8$ (resp. ζ_p) and $R = R_8$ (resp. R_p).

7. Proof of the Second Supplement

Put $\zeta = \zeta_8$, a primitive eighth root of unity and $R = R_8 = \mathbb{Z}[\zeta_8]$. We have:

$$0 = \zeta^8 - 1 = (\zeta^4 + 1)(\zeta^4 - 1).$$

Since $\zeta^4 \neq 1$ (primitivity), we must have $\zeta^4 + 1 = 0$. Multiplying by ζ^{-2} we get

$$\zeta^2 + \zeta^{-2} = 0.$$

So

$$(\zeta + \zeta^{-1})^2 = \zeta^2 + \zeta^{-2} + 2 = 2.$$

Putting $\tau = \zeta + \zeta^{-1}$, we have $\tau^2 = 2$. Now we calculate

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

The "$\equiv$" is by Euler's relation. Multiplying through by τ, we get:

$$\tau^p \equiv \left(\frac{2}{p}\right)\tau \pmod{p}. \tag{23}$$

LEMMA 4.20. *("Schoolboy binomial theorem")*
Let R be a commutative ring, p a prime number and $x, y \in R$. We have

$$(x+y)^p \equiv x^p + y^p \pmod{pR}.$$

PROOF. The binomial formula asserts that

$$(x+y)^p = x^p + \binom{p}{1}x^{p-1}y + \binom{p}{2}x^{p-2}y^2 + \ldots + \binom{p}{p-1}x^1y^{p-1} + y^p,$$

where $\binom{p}{i} = \frac{p!}{i!(p-i)!}$. Now suppose $0 < i < p$. Since p is prime, $p!$ is divisible by p and $i!$ and $(p-i)!$, being a product of positive integers all less than p, are not. So each of the binomial coefficients is divisible by p except the first and the last. □

Therefore

$$\tau^p = (\zeta + \zeta^{-1})^p \equiv \zeta^p + \zeta^{-p} \pmod{pR}.$$

Case 1: $p \equiv 1 \pmod 8$. Then $\zeta^p = \zeta$, and hence $\zeta^{-p} = \zeta^{-1}$, so

$$\tau^p \equiv \zeta^p + \zeta^{-p} \equiv \zeta + \zeta^{-1} = \tau \pmod{pR}$$

so (23) becomes

$$\tau \equiv \left(\frac{2}{p}\right)\tau \pmod{pR}.$$

It is tempting to cancel the τ's, but we must be careful: pR need not be a prime ideal of the ring R.[4] But, sneakily, instead of dividing we *multiply* by τ, getting

$$2 \equiv \tau^2 \equiv 2\left(\frac{2}{p}\right) \pmod{pR},$$

which by Lemma 2 means that

$$2 \equiv 2\left(\frac{2}{p}\right) \pmod{p}$$

in the usual sense. Since 2 is a unit in $\mathbb{Z}/p\mathbb{Z}$, dividing both sides by 2 is permissible. We do so, getting the desired conclusion in this case:

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

Case 2: $p \equiv -1 \pmod 8$ is very similar: this time $\zeta^p = \zeta^{-1}$, but still $\tau^p \equiv \zeta^p + \zeta^{-p} \equiv \zeta^{-1} + \zeta = \tau \pmod{pR}$. The remainder of the argument is the same, in particular the conclusion: $\left(\frac{2}{p}\right) \equiv 1 \pmod p$.

[4] In fact, it can be shown *not* to be prime in the case $p \equiv 1 \pmod 8$.

Case 3: $p \equiv 3 \pmod 8$. Now we have

$$\tau^p \equiv \zeta^p + \zeta^{-p} \equiv \zeta^3 + \zeta^{-3} \equiv \zeta^4\zeta^{-1} + \zeta^{-4}\zeta \equiv -(\zeta + \zeta^{-1}) \equiv -\tau \pmod{pR}.$$

Thus we get this time

$$-\tau \equiv \left(\frac{2}{p}\right)\tau \pmod{pR},$$

and again we multiply by τ to get a congruence modulo p and conclude

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

Case 4: $p \equiv 5 \pmod 8$: Exercise (Case 4 is to Case 3 as Case 2 is to Case 1.)

8. Proof of the Quadratic Reciprocity Law Modulo...

The above proof is due in spirit to Euler. It is very ingenious, but how do we adapt it to prove the Quadratic Reciprocity Law: e.g., what should play the role of τ?

Let us now take p to be an odd prime, $\zeta = e^{\frac{2\pi i}{p}}$ to be a primitive pth root of unity, and $R = \mathbb{Z}[\zeta]$. A good start would be to find an explicit element τ of R with $\tau^2 = p$.

This would mean in particular that $\mathbb{Q}(\sqrt{p}) \subset \mathbb{Q}(\zeta)$, which is far from obvious. Indeed, it need not even be quite true. Take $p = 3$: since $\zeta_3 = e^{\frac{2\pi i}{3}} = (\frac{1-\sqrt{-3}}{2})$, the cyclotomic field $\mathbb{Q}(\zeta_3)$ is the same as the imaginary quadratic field $\mathbb{Q}(\sqrt{-3})$. There is an element $\tau \in \mathbb{Z}[\zeta_3]$ with $\tau^2 = -3$ but not one with $\tau^2 = 3$.

But take heart: finding a square root of p in $\mathbb{Q}(\zeta_p)$ isn't exactly what we wanted anyway. Recall that a strange factor of ± 1 according to whether $p \equiv \pm 1 \pmod 4$ is the hallmark of quadratic reiciprocity. So actually we are on the right track.

Now, like a *deus ex machina* comes the **Gauss sum**:[5]

$$\tau := \sum_{t=0}^{p-1} \left(\frac{t}{p}\right)\zeta^t.$$

In other words, we sum up all the pth roots of unity, but we insert ± 1 signs in front of them according to a very particular recipe. This looks a bit like a random walk in the complex plane with p steps of unit length. A probabilist would guess that the magnitude of the complex number τ is roughly $\sqrt{p}$.[6] Well, it is our lucky day:

THEOREM 4.21. *(Gauss)*

$$\tau^2 = (-1)^{\frac{p-1}{2}}p.$$

That is, $|\tau| = \sqrt{p}$ on the nose! The extra factor of $(-1)^{\frac{p-1}{2}}$ is more than welcome, since it appears in the quadratic reciprocity law. In fact, we define $p^* = (-1)^{\frac{p-1}{2}}p$, and then it is entirely straightforward to check the following

[5]We make the convention that from now until the end of the handout, all sums extend over $0 \leq i \leq p-1$.

[6]Much more on this, the **philosophy of almost square root error**, can be found in the analytic number theory part of these notes.

LEMMA 4.22. *The quadratic reciprocity law is equivalent to the fact that for distinct odd primes p and q, we have*

$$\left(\frac{q}{p}\right) = \left(\frac{p^*}{q}\right).$$

PROOF. Exercise. □

Remarkably, we can now push through a proof as in the last section:

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

and suddenly our way is clear: multiply by τ to get

$$\tau^q \equiv \left(\frac{p^*}{q}\right)\tau \pmod{q}. \tag{24}$$

On the other hand, we have

$$\tau^q \equiv \left(\sum_t \left(\frac{t}{p}\right)\zeta^t\right)^q \equiv \sum_t \left(\frac{t}{p}\right)\zeta^{qt} \pmod{q}.$$

Now, since q is prime to p and hence to the order of ζ, the elements ζ^{qt} still run through all distinct pth roots of unity as t runs from 0 to $p-1$. In other words, we can make the change of variable $t \mapsto q^{-1}t$ and then the sum becomes

$$\sum_t \left(\frac{q^{-1}t}{p}\right)\zeta^t = \left(\frac{q^{-1}}{p}\right)\sum_t \left(\frac{t}{p}\right)\zeta^t = \left(\frac{q}{p}\right)\tau.$$

So we win: substituting this into (24) we get

$$\left(\frac{q}{p}\right)\tau \equiv \left(\frac{p^*}{q}\right)\tau \pmod{q},$$

and multiplying through by τ we get an ordinary congruence

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

since p^* is prime to q, we may cancel to get

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

and finally that

$$\left(\frac{q}{p}\right) = \left(\frac{p^*}{q}\right).$$

9. ... the Computation of the Gauss Sum

Of course, it remains to prove Theorem 4.21. We wish to show that if

$$\tau = \sum_t \left(\frac{t}{p}\right)\zeta^t,$$

then

$$\tau^2 = p^* = (-1)^{\frac{p-1}{2}}p.$$

We do this by introducing a slightly more general sum: for any integer a, we define

$$\tau_a := \sum_t \left(\frac{t}{p}\right) \zeta^{at}.$$

If $a \equiv 0 \pmod{p}$, then

$$\tau_a = \sum_t \left(\frac{t}{p}\right) \zeta^{ap} = \sum_t \left(\frac{t}{p}\right).$$

Notice that τ_q came up in the proof of the quadratic reciprocity law and we quickly rewrote it in terms of τ. That argument still works here, to give:

$$\tau_a = \left(\frac{a}{p}\right) \tau.$$

Now we will evaluate the sum $\sum_a \tau_a \tau_{-a}$ in two different ways. First, if $a \neq 0$, then

$$\tau_a \tau_{-a} = \left(\frac{a}{p}\right)\left(\frac{-a}{p}\right) \tau^2 = \left(\frac{-1}{p}\right) \tau^2 = (-1)^{\frac{p-1}{2}} \tau^2.$$

On the other hand

$$\tau_0 = \sum_t \left(\frac{t}{p}\right) \zeta^{0t} = \sum_t \left(\frac{t}{p}\right) = 0,$$

since each nonzero quadratic residue mod p contributes $+1$, each quadratic non-residue contributes -1, and we have an equal number of each. It follows that

$$\sum_a \tau_a \tau_{-a} = (-1)^{\frac{p-1}{2}} (p-1)\tau^2.$$

We also have

$$\tau_a \tau_{-a} = \sum_x \sum_y \left(\frac{x}{p}\right)\left(\frac{y}{p}\right) \zeta^{a(x-y)}.$$

LEMMA 4.23.
a) If $a \equiv 0 \pmod{p}$, then $\sum_t \zeta^{at} = p$;
b) Otherwise $\sum_t \zeta^{at} = 0$.

The proof is easy. So interchanging the summations we get

$$\sum_a \tau_a \tau_{-a} = \sum_x \sum_y \left(\frac{x}{p}\right)\left(\frac{y}{p}\right) \sum_a \zeta^{a(x-y)}.$$

The inner sum is 0 for all $x \neq y$, and the outer sum is 0 when $x = y = 0$. For each of the remaining $p-1$ values of $x = y$, we get a contribution to the sum of p, so

$$\sum_a \tau_a \tau_{-a} = (p-1)p.$$

Equating our two expressions for $\sum_a \tau_a \tau_{-a}$ gives

$$(p-1)p = (-1)^{\frac{p-1}{2}} (p-1)\tau^2,$$

which gives the desired result:

$$\tau^2 = (-1)^{\frac{p-1}{2}} p = p^*.$$

10. Comments

Working through this proof feels a little bit like being an accountant who has been assigned to carefully document a miracle. Nevertheless, every proof of QR I know feels this way, sometimes to an even greater extent. At least in this proof the miracle can be "bottled": there are many fruitful generalizations of Gauss sums, which can be used to prove an amazing variety of results in mathematics, from number theory to partial differential equations (really!).

The proof just given is a modern formulation of Gauss' sixth and last proof, in which his polynomial identities have been replaced by more explicit reference to algebraic integers. In particular I took the proof from the wonderful text of Ireland and Rosen, with only very minor expository modifications. In addition to being no harder than any other proof of QR that I have ever seen, it has other merits: first, it shows that the cyclotomic field $\mathbb{Q}(\zeta_p)$ contains the quadratic field $\mathbb{Q}(\sqrt{p^*})$ – in fact, Galois theory shows that this is the *unique* quadratic field contained in $\mathbb{Q}$ – a fact which comes up again and again in algebraic number theory. Second, the proof can be adapted with relative ease to prove certain generalizations of the quadratic reciprocity law to cubic and biquadratic residues (for this see Ireland and Rosen again). These higher reciprocity laws were much sought by Gauss but found only by his student Eisenstein (not the filmmaker).

Finally, the Gauss sum can be rewritten to look more like the "Gaussians" one studies in continuous mathematics: you are asked in the homework to show that

$$\tau = \sum_t e^{\frac{2\pi i t^2}{p}}.$$

11. The proof of Jacobian Quadratic Reciprocity

We proved quadratic reciprocity for the Legendre symbol but used quadratic reciprocity for the Jacobi symbol. Of course this is a logical gap. One may well wonder why "Jacobian quadratic reciprocity" gets such short shrift. Namely, more than 200 years after Gauss's *Disquisitiones Arithmeticae*, why do we not simply state and prove quadratic reciprocity for the Jacobi symbol? Is it possible to do so?

The answer is *yes*. A direct proof of Jacobian quadratic reciprocity is given, for instance, in [**BC12**, §3.1]. This proof seems (in my rather biased opinion!) to be natural and appealing – in particular it gives some further insight into *why the Jacobi symbol is defined as it is*, whereas in our present treatment it seems to simply be a clever way to compute Legendre symbols without factorization. We give an account of this "Zolotarev-Frobenius" approach to quadratic reciprocity in the next (quite ancillary) chapter.

However, the emphasis on "Gaussian quadratic reciprocity" is justified by the fact that Jacobian Quadratic Reciprocity can be deduced from the "vanilla" version by arguments which are – in comparison with every known proof of Gaussian quadratic reciprocity – rather banal. We give such a deduction now.

LEMMA 4.24. *Let $n \geq 1$, and let $a_1, \ldots, a_n$ be odd integers.*
a) We have

$$\frac{a_1 \cdots a_n - 1}{2} \equiv \sum_{i=1}^n \frac{a_i - 1}{2} \pmod 2.$$

b) We have

$$\frac{(a_1 \cdots a_n)^2 - 1}{8} \equiv \sum_{i=1}^n \frac{a_i^2 - 1}{8} \pmod 2.$$

PROOF. a) We proceed by induction on n. The case $n = 1$ is trivial. The "real base case" is $n = 2$, which we establish by direct calculation:

$$\frac{a_1 a_2 - 1}{2} - \left(\frac{a_1 - 1}{2} + \frac{a_2 - 1}{2}\right) = \frac{(a_1 - 1)(a_2 - 1)}{2} \equiv 0 \pmod 2$$

since $a_1 - 1$ and $a_2 - 1$ are both even. Now we suppose the result holds for $n - 1 \geq 1$ and establish it for n. We have

$$\frac{a_1 \cdots a_n - 1}{2} = \frac{(a_1 \cdots a_{n-1})a_n - 1}{2} \equiv \frac{a_1 \cdots a_{n-1} - 1}{2} + \frac{a_n - 1}{2} \pmod 2,$$

and by induction the right hand side is congruent modulo 2 to

$$\left(\sum_{i=1}^{n-1} \frac{a_i - 1}{2}\right) + \frac{a_n - 1}{2} \equiv \sum_{i=1}^n \frac{a_i - 1}{2} \pmod 2.$$

b) This is quite similar: $n = 1$ is trivial, and the crucial case $n = 2$ is a calculation:

$$\frac{(a_1 a_2)^2 - 1}{8} - \left(\frac{a_1^2 - 1}{8} + \frac{a_2^2 - 1}{8}\right) = \frac{(a_1^2 - 1)(a_2^2 - 1)}{8} \equiv 0 \pmod 2$$

since $a_1^2 - 1$ and $a_2^2 - 2$ are both divisible by 8. Now suppose the result holds for $n - 1 \geq 1$ and establish it for n. We have

$$\frac{(a_1 \cdots a_n)^2 - 1}{8} = \frac{(a_1 \cdots a_{n-1})a_n)^2 - 1}{8} \equiv \frac{(a_1 \cdots a_{n-1})^2 - 1}{8} + \frac{a_n^2 - 1}{8} \pmod 2,$$

and by induction the right hand side is congruent modulo 2 to

$$\left(\sum_{i=1}^{n-1} \frac{a_i^2 - 1}{8}\right) + \frac{a_n^2 - 1}{8} \equiv \sum_{i=1}^n \frac{a_i^2 - 1}{8} \pmod 2. \qquad \square$$

Now we prove Jacobian quadratic reciprocity. Let a and b be odd positive integers, with prime factorizations

$$a = p_1 \cdots p_m, \ b = q_1 \cdots q_n.$$

Step 1: Using the first supplement to Quadratic Reciprocity and Lemma 4.24a), we get

$$\left(\frac{-1}{b}\right) = \prod_{j=1}^n \left(\frac{-1}{q_j}\right) = \prod_{j=1}^n (-1)^{\frac{q_j - 1}{2}}$$

$$= (-1)^{\sum_{j=1}^n \frac{q_j - 1}{2}} = (-1)^{q_1 \cdots q_n - 1} 2 = (-1)^{\frac{b-1}{2}}.$$

Step 2: Similarly, using the second supplement to Quadratic Reciprocity and Lemma 4.24b), we get

$$\left(\frac{2}{b}\right) = \prod_{j=1}^n \left(\frac{2}{q_j}\right) = \prod_{j=1}^n (-1)^{\frac{q_j^2 - 1}{8}}$$

$$= (-1)^{\sum_{j=1}^n \frac{q_j^2-1}{8}} = (-1)^{\frac{(q_1\cdots q_n)^2-1}{8}} = (-1)^{\frac{b^2-1}{8}}.$$

Step 3: Put

$$S \coloneqq \sum_{1\le i\le m,\ 1\le j\le n} \frac{p_i-1}{2}\frac{q_j-1}{2}.$$

Then using Lemma 4.24a) twice, we get

$$S = \sum_{1\le j\le n}\left(\sum_{1\le i\le m}\frac{p_i-1}{2}\right)\frac{q_j-1}{2} \equiv \sum_{1\le j\le n}\left(\frac{a-1}{2}\right)\frac{q_j-1}{2}$$
$$\equiv \left(\frac{a-1}{2}\right)\sum_{1\le j\le n}\frac{q_j-1}{2} \equiv \frac{a-1}{2}\frac{b-1}{2} \pmod 2,$$

and thus

$$\left(\frac{a}{b}\right)\left(\frac{b}{a}\right) = \prod_{1\le i\le m,\ 1\le j\le n}\left(\frac{p_i}{q_j}\right)\left(\frac{q_j}{p_i}\right)$$
$$= \prod_{1\le i\le m,\ 1\le j\le n}(-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}} = (-1)^S = (-1)^{\frac{a-1}{2}\frac{b-1}{2}}.$$

CHAPTER 5

More Quadratic Reciprocity: from Zolotarev to Duke-Hopkins

1. Quadratic Reciprocity in a Finite Quotient Domain

An **FQ-domain** is a commutative domain R with the property that for every $x \in R^\bullet$, $R/(x)$ is finite. Evidently any field is an FQ-domain, and conversely any finite domain is a field. Here we will be interested in FQ-domains which are not fields.

Let R be a finite quotient domain, and let $x \in R^\bullet$. We define the **norm** $|x| := \#R/(x)$, and put $|0| = 0$. We claim that for $x, y \in R^\bullet$, $|xy| = |x||y|$: indeed, this follows from the short exact sequence

$$0 \to \frac{(x)}{(xy)} \to R/(xy) \to R/(x) \to 0$$

and the fact that multiplication by x gives an isomorphism $R/(y) \xrightarrow{\sim} (x)/(xy)$.

Example: If R is a domain with additive group isomorphic to $\mathbb{Z}^n$ for some $n \geq 1$, then R is an FQ-domain. In particular, for any number field K, the ring of integers $\mathbb{Z}_K$ is an FQ-domain.

Example: Let $C_{/\mathbb{F}_q}$ be an integral, normal affine algebraic curve. Then the coordinate ring $\mathbb{F}_q[C]$ is an FQ-domain. In particular, the univariate polynomial ring $\mathbb{F}_q[t]$ is an FQ-domain.

As usual, a **prime** (or **prime element**) in R is a nonzero element p such that pR is a prime ideal. In a finite quotient domain, if p is a prime element, then $R/(p)$ is a finite field: in particular the ideal (p) is maximal.

Remark: A theorem of Kaplansky states that a domain R is a UFD iff every nonzero prime ideal $\mathfrak{p}$ of R contains a prime element. It is also known that a domain is a PID iff every prime ideal is principal. From these two results it follows that an FQ-domain is a UFD iff it is a PID.

We say that a nonzero, nonunit $x \in R$ is **factorable** if there exist primes $p_1, \cdot, p_r$ such that $x = p_1 \cdots p_r$. An FQ-domain is a UFD iff every nonzero nonunit is factorable, but even if R is not a UFD, then factorization of an element into *primes* (as opposed to merely irreducibles) is necessarily unique up to associates.

We say $x \in R^\bullet$ is **odd** if $|x|$ is an odd integer. Note that if $2 \in R^\times$ then every nonzero x is odd, whereas if $2 = 0$ in R then no element of x is odd.

Let I be a nonzero ideal of the FQ-domain R. Then I contains a nonzero element x, we have a natural surjection $R/(x) \to R/I$, and since $R/(x)$ is finite, so is R/I. We may therefore extend the "norm map" to nonzero ideals by $|I| = \#R/I$ and also put $|(0)| = 0$. Note that this generalizes the previous norm map in that for all $x \in R^\bullet$ we have $|(x)| = |x|$. As above, we say an ideal I is **odd** if $|I|$ is odd. A nonzero proper ideal $\mathfrak{b}$ of R is **factorable** if there exist prime ideals $\mathfrak{p}_1, \ldots, \mathfrak{p}_r$ of R such that $\mathfrak{b} = \prod_{i=1}^r \mathfrak{p}_i$. Note that if the element b is factorable, then so is the principal ideal (b), but in general the converse does not hold. Because of this we say that an element b is **I-factorable** if the ideal (p) factors into a product of prime ideals.

For $a \in R$ and an odd prime ideal $\mathfrak{p}$, we define the **Legendre symbol** $\left(\frac{a}{\mathfrak{p}}\right)$: it is 0 if $a \in \mathfrak{p}$, 1 if $a \notin \mathfrak{p}$ and $a \equiv x^2 \pmod{\mathfrak{p}}$, and -1 if $a \notin \mathfrak{p}$ and $a \not\equiv x^2 \pmod{\mathfrak{p}}$. For an odd factorable ideal $\mathfrak{b} = \mathfrak{p}_1 \cdots \mathfrak{p}_r$ of R we define the **Jacobi symbol**

$$\left(\frac{a}{\mathfrak{b}}\right) = \prod_{i=1}^r \left(\frac{a}{\mathfrak{b}}\right).$$

Let $\mathfrak{r}$ be a ring, and let $a \in \mathfrak{r}^\times$. Then the map $m_a : \mathfrak{r} \to \mathfrak{r}$ by $x \mapsto xa$ is a bijection (its inverse is $\bullet a^{-1}$).

Now suppose moreover $\mathfrak{r}$ is finite, of order n, so upon choosing a bijection of R with $\{1, \ldots, n\}$, we may identify $\bullet a$ with an element of the symmetric group S_n, and in particular $\bullet a$ has a well-defined **sign** $[\frac{a}{\mathfrak{r}}] \in \{\pm 1\}$ The sign map $\epsilon : S_n \to \{\pm 1\}$ is a homorphism into a commutative group, so for all $\sigma, \tau \in S_n$, $\epsilon(\tau\sigma\tau^{-1}) = \epsilon(\sigma)$. In particular $[\frac{a}{\mathfrak{r}}]$ is independent of the choice of bijection $\mathfrak{r} \xrightarrow{\sim} \{1, \ldots, n\}$.)

These two constructions are related, as follows: let R be an FQ-domain, $\mathfrak{b} = \mathfrak{p}_1 \cdots \mathfrak{p}_r$ a nonzero, proper factorable odd ideal of R, and a an element of R which is **relatively prime** to $\mathfrak{b}$ in the sense that $(a) + \mathfrak{b} = R$. Then (the image of) a is a unit in the finite ring $R/(\mathfrak{b})$ so that $\left[\frac{a}{R/\mathfrak{b}}\right]$ is well-defined.

THEOREM 5.1. *(Zolotarev's First Lemma) Let R be an abstract number ring, $\mathfrak{b}$ an odd factorable ideal of R, and a an element of R which is relatively prime to $\mathfrak{b}$. Then*

$$\left[\frac{a}{R/\mathfrak{b}}\right] = \left(\frac{a}{\mathfrak{b}}\right).$$

PROOF. ... □

1.1. The Zolotarev Symbol.

Now let a and b be two relatively prime odd elements of the FQ-domain R. We will define three permutations ζ, α, β of the finite set $R/(a) \times R/(b)$.

THEOREM 5.2. *(Zolotarev's Second Lemma)*

a) We have $\epsilon(\alpha) = \left[\frac{a}{R/(b)}\right] = \left(\frac{a}{b}\right)$.

b) We have $\epsilon(\beta) = \left[\frac{b}{R/(a)}\right] = \left(\frac{b}{a}\right)$.

c) We have $\epsilon(\zeta) = \epsilon(\alpha) \cdot \epsilon(\beta)$.

For coprime odd $a, b \in R$, we define the **Zolotarev symbol**

$$Z(a,b) = \epsilon(\zeta) \in \{\pm 1\}.$$

THEOREM 5.3. *(Abstract Quadratic Reciprocity) Let a, b be relatively prime, odd I-factorable elements of an FQ-domain R. Then*

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

2. The Kronecker Symbol

The **Jacobi symbol** $(\frac{a}{n})$ is an extension of the Legendre symbol $(\frac{a}{p})$ which is defined for any positive odd integer n by $(\frac{a}{1}) = 1$ for all $a \in \mathbb{Z}$; if $n = \prod_{i=1}^r p_i$, then

$$\left(\frac{a}{n}\right) = \prod_{i=1}^r \left(\frac{a}{p_i}\right).$$

For an integer a, define

$$\left(\frac{a}{2}\right) = \left\{ \begin{array}{ccc} 0 & a \equiv 0 & (\mathrm{mod}\ 2) \\ 1 & a \equiv 1,7 & (\mathrm{mod}\ 8) \\ -1 & a \equiv 3,5 & (\mathrm{mod}\ 8) \end{array} \right\},$$

$$\left(\frac{a}{-1}\right) = \left\{ \begin{array}{cc} 0 & a = 0 \\ 1 & a > 0 \\ -1 & a < 0 \end{array} \right\},$$

$$\left(\frac{a}{0}\right) = \left\{ \begin{array}{cc} 0 & a \neq 1 \\ 1 & a = 1 \end{array} \right\}.$$

With these additional rules there is a unique extension of the Jacobi symbol to a symbol $(\frac{n}{a})$ defined for any $n, a \in \mathbb{Z}$ such that for all integers n, a, b, we have $(\frac{n}{ab}) = (\frac{n}{a})(\frac{n}{b})$. One also has $(\frac{ab}{n}) = (\frac{a}{n})(\frac{b}{n})$, i.e., the symbol is **bi-multiplicative**. This extension of the Jacobi symbol is known as the **Kronecker symbol**.

When n is **not** odd and positive, some authors (e.g. [**DH05**]) define $(\frac{a}{n})$ only when $a \equiv 0, 1 \pmod 4$. It is not worth our time to discuss these two conventions, but we note that all of our results involve only this"restricted" Kronecker symbol.

For odd $n \in \mathbb{Z}^+$, define $n^* = (-1)^{\frac{n-1}{2}} n$. **Full quadratic reciprocity** – i.e., the usual QR law together with its First and Second Supplements – is equivalent to one elegant identity: for $a \in \mathbb{Z}$ and an odd positive $n \in \mathbb{Z}$,

$$\left(\frac{a}{n}\right) = \left(\frac{n^*}{a}\right). \tag{25}$$

3. The Duke-Hopkins Reciprocity Law

Let G be a finite commutative group (written multiplicatively) of order n. We define an action of $(\mathbb{Z}/n\mathbb{Z})^\times$ on G, by

$$(a \mod n) \bullet g := g^a.$$

By Lagrange's Theorem, $g^n = 1$, so that $g^a = g^{a'}$ if $a \equiv a' \pmod n$ and $a\bullet$ is well defined. It is immediate that each $a\bullet$ gives a homomorphism from G to G;

moreover, since $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, there exists $b \in (\mathbb{Z}/n\mathbb{Z})^\times$ such that $ab \equiv 1 \pmod{n}$, and then $a\bullet \circ b \bullet = b \bullet \circ a\bullet = \mathrm{Id}_G$, so that each $a\bullet$ is an automorphism of G.

As for any group action on a set, this determines a homomorphism from $(\mathbb{Z}/n\mathbb{Z})^\times$ to the group $\mathrm{Sym}(G)$ of permutations of G, the latter group being isomorphic to S_n, the symmetric group on n elements. Recall that there is a unique homomorphism from S_n to the cyclic group Z_2 given by the **sign** of the permutation. Therefore we have a composite homomorphism

$$(\mathbb{Z}/n\mathbb{Z})^\times \to \mathrm{Sym}(G) \to Z_2$$

which we will denote by

$$a \pmod{n} \mapsto \left(\frac{a}{G}\right).$$

Example 2.1 (Zolotarev): Let p be an odd prime and $G = Z_p$ is the cyclic group of order p. The mapping $(\mathbb{Z}/p\mathbb{Z})^\times \to Z_2$ given by $a \mapsto \left(\frac{a}{Z_p}\right)$ is nothing else than the usual Legendre symbol $a \mapsto (\frac{a}{p})$. Indeed, the group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic of even order, so admits a unique surjective homomorphism to the group $Z_2 = \{\pm 1\}$: if g is a primitive root mod p, we send g to -1 and hence every odd power of g to -1 and every even power of g to $+1$. This precisely describes the Legendre symbol $a \mapsto (\frac{a}{p})$. Thus it suffices to see that for some $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ we have $\left(\frac{a}{Z_p}\right) = -1$, i.e., the sign of the permutation $n \in Z_p \mapsto n^a$ is -1. To see this, switch to additive notation, viewing Z_p as the isomorphic group $(\mathbb{Z}/p\mathbb{Z}, +)$; the action in question is now just multiplication by a nonzero element a. If g is a primitive root modulo p, multiplication by g fixes 0 and cyclically permutes all $p-1$ nonzero elements, so is a cycle of even order and hence an odd permutation: thus $(\frac{g}{Z_p}) = -1$.

The next result shows that the symbol $(\frac{a}{G})$ is also bi-multiplicative.

PROPOSITION 5.4. *For $i = 1, 2$ let G_i be a finite commutative group of order n_i and $a \in (\mathbb{Z}/n_1n_2\mathbb{Z})^\times$. Then*

$$\left(\frac{a}{G_1 \times G_2}\right) = \left(\frac{a \pmod{n_1}}{G_1}\right)\left(\frac{a \pmod{n_2}}{G_2}\right).$$

PROOF. If $a \in (\mathbb{Z}/n_1n_2\mathbb{Z})^\times$, then

$$a \bullet (g_1, g_2) = (g_1^a, g_2^a) = (g_1^{a \pmod{n_1}}, g_2^{a \pmod{n_2}}).$$

After identifying G_1 (resp. G_2) with the subset $G_1 \times e_{G_2}$ (resp. $e_{G_1} \times G_2$) of $G_1 \times G_2$, the permutation that a induces on $G_1 \times G_2$ is the product of the permutation that $a \pmod{n}_1$ induces on G_1 with the permutation that $a \pmod{n}_2$ induces on G_2. □

Let us now consider the action of -1 on $\mathrm{Sym}(G)$. Let r_1 be the number of fixed points of $-1\bullet$. More concretely, $-1 \bullet g = g^{-1} = g$ iff g has order 1 or 2. Note that $r_1 \geq 1$ because of the identity element. The $n - r_1$ other elements of G are all distinct from their multiplicative inverses, so there exists a positive integer r_2 such that $n - r_1 = 2r_2$.

Definition: We put $G^* = (-1)^{r_2}|G|^{r_1} = (-1)^{r_2}n^{r_1}$.

LEMMA 5.5. *For any finite commutative group G, we have $G^* \equiv 0$ or $1 \pmod{4}$.*

PROOF. Let $n = |G|$. If n is odd, then by Lagrange the only g with $g^{-1} = g$ is the identity, so that $r_1 = 1$ and $r_2 = \frac{n-1}{2}$. In this case $G^* = |G|^* = (-1)^{\frac{n-1}{2}} n \equiv 1 \pmod 4$. If n is even, then $n - r_1 = 2r_2 \equiv 0 \pmod 2$, so r_1 is even and hence is at least 2, so $G^* = (-1)^{r_2} n^{r_1} \equiv 0 \pmod 4$. □

So the Kronecker symbol $\left(\frac{G^*}{a}\right)$ is always defined (even in the "restricted" sense).

THEOREM 5.6. *(Duke-Hopkins Reciprocity Law) For a finite commutative group G and an integer a, we have*

$$\left(\frac{a}{G}\right) = \left(\frac{G^*}{a}\right).$$

The proof will be given in the next section.

COROLLARY 5.7. *a) Suppose G has odd order n. Then for any $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, we have*

$$\left(\frac{a}{G}\right) = \left(\frac{n^*}{a}\right).$$

b) Taking $G = Z_n$ we recover (34).
c) We have $(\frac{a}{G}) = 1$ for all $a \in (\mathbb{Z}/n\mathbb{Z})^\times$ iff n is a square.

Proof of Corollary 5.7: In the proof of Lemma 5.5 we saw that $G^* = n^*$; part a) then follows immediately from the reciprocity law. By part a), the symbol $(\frac{a}{G})$ can be computed using any group of order n, so factor n into a product $p_1 \cdots p_r$ of not necessarily distinct primes and apply Example 2.1: we get $(\frac{a}{G}) = \prod_{i=1}^r (\frac{a}{p_i}) = (\frac{a}{n})$. This gives part b). Finally, using the Chinese Remainder Theorem it is easy to see that there is some a such that $(\frac{a}{n}) = -1$ iff n is not a square.

4. The Proof

Enumerate the elements of G as $g_1, \ldots, g_n$ and the characters of G as $\chi_1, \ldots, \chi_n$. Let M be the $n \times n$ matrix whose (i, j) entry is $\chi_i(g_j)$.

Since any character $\chi \in X(G)$ has values on the unit circle in $\mathbb{C}$, we have $\chi^{-1} = \overline{\chi}$. Therefore the number r_1 of fixed points of -1 on G is the same as the number of characters χ such that $\overline{\chi} = \chi$, i.e., real-valued characters. Thus the effect of complex conjugation on the character matrix M is to fix each row corresponding to a real-valued character and to otherwise swap the ith row with the jth row where $\chi_j = \overline{\chi_i}$. In all r_2 pairs of rows get swapped, so

$$\det(\overline{M}) = \det(M) \cdot (-1)^{r_2}.$$

Moreover, with $M^* = (\overline{M})^t$, we have

$$MM^* = nI_n,$$

so that

$$\det(M) \det(\overline{M}) = n^n,$$

so

$$\det(M)^2 = (-1)^{r_2} n^n = (-1)^{r_2} n^{r_1} n^{2r_2} = \ell^2 G^*, \tag{26}$$

where $\ell = n^{r_2}$. (In particular $\det(M)^2$ is a positive integer. Note that $\det(M)$ itself lies in $\mathbb{Q}(\sqrt{G^*})$, and is not rational if n is odd.) So for any $a \in \mathbb{Z}$, we have

$$\left(\frac{\det(M)^2}{a}\right) = \left(\frac{G^*}{a}\right). \tag{27}$$

The character matrix M has values in the cyclotomic field $\mathbb{Q}(\zeta_n)$, which is a Galois extension of $\mathbb{Q}$, with Galois group isomorphic to (what a concidence!) $(\mathbb{Z}/n\mathbb{Z})^\times$, an explicit isomorphism being given by making $a \in (\mathbb{Z}/n\mathbb{Z})^\times$ correspond to the unique automorphism σ_a of $\mathbb{Q}(\zeta_n)$ satisfying $\sigma_a(\zeta_n) = \zeta_n^a$. (All of this is elementary Galois theory except for the more number-theoretic fact that the cyclotomic polynomial Φ_n is irreducible over $\mathbb{Q}$.) In particular the group $(\mathbb{Z}/n\mathbb{Z})^\times$ also acts by permutations on the character group $X(G)$, and indeed in exactly the same way it acts on G:

$$\forall\ g \in G,\ (a \bullet \chi)(g) = \chi(g^a) = (\chi(g))^a = \chi^a(g),$$

so $a \bullet \chi = \chi^a$. This has the following beautiful consequence:

For $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, applying the Galois automorphism σ_a to the character matrix M induces a permutation of the rows which is "the same" as the permutation $\bullet a$ of G. In particular the signs are the same, so

$$\det(\sigma_a M) = \det(M) \cdot \left(\frac{a}{G}\right). \tag{28}$$

Combining (40) and (28), we get that for all $a \in (\mathbb{Z}/n\mathbb{Z})^\times$,

$$\sigma_a(\sqrt{G^*}) = \left(\frac{a}{G}\right)\sqrt{G^*}.$$

Now, by the multiplicativity on both sides it is enough to prove Theorem 5.6 when $a = p$ is a prime not dividing n and when $a = -1$.

PROPOSITION 5.8. *Let p be a prime not dividing n. TFAE:*
a) $\sigma_p(\sqrt{G^*}) = \sqrt{G^*}$.
b) p splits in $\mathbb{Q}(\sqrt{G^*})$.
c) $(\frac{G^*}{p}) = 1$.

The proof of this – a standard result in algebraic number theory – is omitted for now.

We deduce that

$$\left(\frac{G^*}{p}\right) = \left(\frac{p}{G}\right).$$

Finally, when $a = -1$, σ_{-1} is simply complex conjugation, so

$$\left(\frac{-1}{G}\right)\sqrt{G^*} = \sigma_{-1}(\sqrt{G^*}) = \left\{ \begin{array}{ll} \sqrt{G^*} & G^* > 0 \\ -\sqrt{G^*} & G^* < 0 \end{array} \right\} = \left(\frac{G^*}{-1}\right)\sqrt{G^*},$$

so

$$\left(\frac{-1}{G}\right) = \left(\frac{G^*}{-1}\right).$$

This completes the proof of Theorem 5.6.

5. In Fact...

...the "real" Duke-Hopkins reciprocity law is an assertion about a group G of order n which is not necessarily commutative. In this case, the map $g \mapsto g^a$ need not be an automorphism of G, so a more sophisticated approach is needed. Rather, one considers the action of $(\mathbb{Z}/n\mathbb{Z})^\times$ on the **conjugacy classes** $\{C_1, \dots, C_m\}$ of G: if $g = xhx^{-1}$ then $g^a = xh^a x^{-1}$, so this makes sense. We further define r_1 to be the number of "real" conjugacy classes $C = C^{-1}$ – and assume that in our labelling $C_1, \dots, C_{r_1}$ are all real – and define r_2 by the equation $m = r_1 + 2r_2$. Then in place of our G^* (notation which is not used in [**DH05**]), one has the **discriminant**

$$d(G) = (-1)^{r_2} n^{r_1} \prod_{j=1}^{r_1} |C_j|^{-1}.$$

The Duke-Hopkins reciprocity law asserts that for $a \in (\mathbb{Z}/n\mathbb{Z})^\times$,

$$\left(\frac{a}{G}\right) = \left(\frac{d(G)}{a}\right).$$

The proof is very similar, except the group $X(G)$ of one-dimensional characters gets replaced by the set $\{\chi_1, \dots, \chi_m\}$ of characters (i.e., trace functions) of the irreducible complex representations of G. Perhaps surprisingly, the only part of the proof which looks truly deeper is the claim that $d(G) \equiv 0, 1 \pmod 4$ which is required, according to the conventions of [**DH05**], for the Kronecker symbol $(\frac{d(G)}{a})$ can be defined. Duke and Hopkins suggest this as an analogue of Stickelberger's theorem in algebraic number theory which asserts that the discriminant of any number field is an integer which is 0 or 1 modulo 4; moreover they adapt a 1928 proof of that theorem due to Issai Schur.

CHAPTER 6

The Mordell Equation

1. The Coprime Powers Trick in $\mathbb{Z}$

We have by now seen several ways in which the fundamental theorem of arithmetic can be used to solve Diophantine equations, and that suitably generalized, these techniques often apply to more general unique factorization domains.

We will now consider another such technique, the **coprime powers trick**. In the interest of linear exposition, we present the technique first and then give an application. However, the reader might prefer to skip ahead and see how it is used.

PROPOSITION 6.1. *(Coprime Powers Trick, v. 1)*
Let $n \in \mathbb{Z}^+$, let $x, y, z \in \mathbb{Z}$ be such that $\gcd(x, y) = 1$ and $xy = z^n$.
a) There exist $a, b \in \mathbb{Z}$ such that $x = \pm a^n,\ y = \pm b^n$.
b) If n is odd, then there exist $a, b \in \mathbb{Z}$ such that $x = a^n,\ y = b^n$.

PROOF. If $x, y \in \mathbb{Z}$, then $x = \pm y$ iff $\mathrm{ord}_p(x) = \mathrm{ord}_p(y)$ for all prime numbers p. We exploit this as follows: for any prime p, take ord_p of both sides of $xy = z^n$ to get

$$\mathrm{ord}_p(x) + \mathrm{ord}_p(y) = n\,\mathrm{ord}_p(z).$$

Since x and y are relatively prime, at least one of $\mathrm{ord}_p(x),\ ord_p(y)$ is equal to 0, and therefore they are both divisible by n. Now define $a, b \in \mathbb{Z}^+$ as follows:

$$a = \prod_p p^{\frac{\mathrm{ord}_p(x)}{n}},\ b = \prod_p p^{\frac{\mathrm{ord}_p(y)}{n}}.$$

Then for all primes p, $\mathrm{ord}_p(a^n) = n\,\mathrm{ord}_p(a) = \mathrm{ord}_p(x)$ and $\mathrm{ord}_p(b^n) = n\,\mathrm{ord}_p(b) = \mathrm{ord}_p(y)$. We conclude $x = \pm a^n$, $y = \pm b^n$, establishing part a). Part b) follows upon noticing that if n is odd, $(-1)^n = -1$, so we may write $x = (\pm a)^n$, $y = (\pm b)^n$. □

1.1. An application.

THEOREM 6.2. *The only integral solutions to*

$$y^2 - y = x^3 \tag{29}$$

are $(0, 0)$ and $(0, 1)$.

PROOF. Suppose $(x, y) \in \mathbb{Z}^2$ satisfy equation (29), i.e., $y(y-1) = x^3$. As for any two consecutive integers, y and $y - 1$ are relatively prime. We can therefore apply Proposition 6.1b) to conclude that there exist $a, b \in \mathbb{Z}$ such that

$$y = a^3,\ y - 1 = b^3.$$

This gives

$$1 = y - (y-1) = a^3 - b^3 = (a-b)(a^2 + ab + b^2),$$

and the only way this can happen is for

$$a - b = a^2 + ab + b^2 = \pm 1.$$

Suppose first that $a - b = 1$, so $b = a - 1$; then

$$1 = a^2 + ab + b^2 = a^2 + a(a-1) + (a-1)^2 = 3a^2 - 3a + 1,$$

or

$$3a^2 - 3a = 0.$$

The solutions of this quadratic are $a = 0$ and $a = 1$. If $a = 0$, then $y = a^3 = 0$, and $x^3 = 0^2 - 0 = 0$: we get the solution $(x, y) = (0, 0)$ to (29). If $a = 1$, then $y = 1$ and $x^3 = 1^2 - 1 = 0$: we get the solution $(x, y) = (0, 1)$.
Next suppose that $a - b = -1$, so $b = a + 1$; then

$$-1 = a^2 + ab + b^2 = a^2 + a(a+1) + (a+1)^2 = 3a^2 + 3a + 1,$$

or

$$3a^2 + 3a + 2 = 0,$$

a quadratic equation with discriminant $3^2 - 4 \cdot 3 \cdot 2 = -13 < 0$; thus there are no real solutions. □

2. The Mordell Equation

We now turn to a family of Diophantine equations which has received persistent attention over the centuries and remains of interest to this day. Namely, fix an integer k and consider

$$y^2 + k = x^3. \tag{30}$$

We wish to find all integral solutions. If $k = 0$ we get the "degenerate" equation $y^2 = x^3$. A moment's thought shows that this equation has solution set $\{(x, y) = (a^2, a^3) \mid a \in \mathbb{N}\}$. In particular there are infinitely many solutions. The great Philadelphian mathematician Louis J. Mordell showed that conversely, for each nonzero k, (30) has only finitely many integer solutions. Because of this and other results over the course of his long career, (30) is often called the **Mordell Equation**, despite the fact that other distinguished mathematicians also worked on it. In particular, the case of $k = -2$ was considered by Claude-Gaspar Bachet and Fermat in the seventeenth century, and the following result is attributed to Fermat.

THEOREM 6.3. *(Fermat) The only integral solutions to*

$$y^2 + 2 = x^3 \tag{31}$$

are $(3, 5)$ *and* $(3, -5)$.

PROOF. We wish to argue similarly to the previous result, but here the only factorization in sight takes place over the quadratic ring $\mathbb{Z}[\sqrt{-2}]$, namely:

$$x^3 = (y + \sqrt{-2})(y - \sqrt{-2}).$$

Looking back at the previous argument, it seems that what we would like to say is that there are elements $\alpha = a + b\sqrt{-2}, \beta = c + d\sqrt{-2} \in \mathbb{Z}[\sqrt{-2}]$ such that

$$y + \sqrt{-2} = \alpha^3, \ y - \sqrt{-2} = \beta^3.$$

The justification for this will be a version of the coprime powers trick in the ring $\mathbb{Z}[\sqrt{-2}]$, but let us assume it just for a moment and see what comes of it.

By expanding out α^3 we get

$$y + \sqrt{-2} = (a + b\sqrt{-2})^3 = (a^3 - 6ab^2) + (3a^2b - 2b^3)\sqrt{-2},$$

and this means that

$$y = a^3 - 6ab^2 = a(a^3 - 6b^2),$$
$$1 = 3a^2b - 2b^3 = b(3a^2 - 2b^2).$$

Again this very much limits our options: we must have

$$b = 3a^2 - 2b^2 = 1$$

or

$$b = 3a^2 - 2b^2 = -1.$$

Taking the first option – $b = 1$ – gives $3a^2 = 2b^2 + 1 = 3$, so $a = \pm 1$. Taking $(a, b) = (1, 1)$ leads to $y = 1(1^3 - 6 \cdot 1^2) = -5$, so $x^3 = y^2 + 2 = 5^2 + 2 = 27$, so $x = 3$: we get the solution $(x, y) = (3, 5)$. Taking $(a, b) = (-1, 1)$ leads to $y = -1((-1)^3 - 6 \cdot 1^2) = 7$, so $x^3 = y^2 + 2 = 7^2 + 2 = 51$, which has no integral solutions since 51 is not a perfect cube.

The second option – $b = -1$ – gives $3a^2 = 2b^2 + 1 = 3$, so again $a = \pm 1$. Taking $(a, b) = (1, -1)$ leads to $y = 1(1^3 - 6 \cdot (-1)^2) = -5$, and as above we get $x = 3$ and the solution $(x, y) = (3, -5)$. Taking $(a, b) = (-1, -1)$ leads to $y = -1((-1)^3 - 6 \cdot (-1)^2) = 7$, which as above yields no solution. □

The time has come to justify our assumption that there exist elements α, β as above. The justification is in two parts: first, we need a version of the coprime powers trick that applies to the domain $\mathbb{Z}[\sqrt{-2}]$; and second we need to verify that the hypotheses are justified in our particular case: in particular, that the elements $y \pm \sqrt{-2}$ of $\mathbb{Z}[\sqrt{-2}]$ are indeed coprime!

3. The Coprime Powers Trick in a UFD

3.1. Ord functions and coprime powers.

Let R be a UFD and $x, y \in R$. We say x, y are **coprime** if $z \mid x$, $\mid z \mid y$ implies $z \in R^\times$: equivalently, there is no prime element which divides both of them.

Proposition 6.4. *(Coprime powers trick, v. 2) Let R be a UFD, $n \in \mathbb{Z}^+$, and let $x, y, z \in R$ be coprime elements such that $xy = z^n$.*
a) There exist $\alpha, \beta \in R$ and units $u, v \in R^\times$ such that

$$x = u\alpha^n, \ y = v\beta^n.$$

b) If every unit in R is an nth power, then there exist $\alpha, \beta \in R$ such that

$$x = \alpha^n, y = \beta^n.$$

In other words, if in a UFD the product of two relatively prime elements is a perfect nth power, then each of them is a perfect nth power, up to a unit.

Before giving the proof, we set up a more general notion of "ord functions". We work in the context of an integral domain R which satisfies the ascending chain condition on principal ideals (ACCP). In plainer terms we assume that there is no

infinite sequence $\{x_i\}_{i=1}^{\infty}$ of elements of R such that x_{i+1} properly divides[1] x_i for all i. This is a very mild condition: it is satisfied by any Noetherian ring and by any UFD: c.f. [Factorization in Integral Domains].

Now let π be a nonzero prime element of R, and let $x \in R \setminus \{0\}$. (ACCP) ensures that there exists a largest non-negative integer n such that $\pi^n \mid x$, for otherwise $\pi^n \mid x$ for all n and $\{\frac{x}{\pi^n}\}$ is an infinite sequence in which each element properly divides the previous one. We put $\mathrm{ord}_\pi(x)$ to be this largest integer n. In other words, $\mathrm{ord}_\pi(x) = n$ iff $\pi^n \mid x$ and $\pi^{n+1} \nmid x$. We formally set $\mathrm{ord}_\pi(0) = +\infty$, and we extend ord_π to a function on the fraction field K of R by multiplicativity:

$$\mathrm{ord}_\pi\left(\frac{x}{y}\right) := \mathrm{ord}_\pi(x) - \mathrm{ord}_\pi(y).$$

This generalizes the functions ord_p on $\mathbb{Z}$ and $\mathbb{Q}$, and the same properties hold.

PROPOSITION 6.5. *Let R be an (ACCP) domain with fraction field K. Let π be a nonzero prime element of R and $x, y \in K \setminus \{0\}$. Then:*
a) $\mathrm{ord}_\pi(xy) = \mathrm{ord}_\pi(x) + \mathrm{ord}_\pi(y)$.
b) $\mathrm{ord}_\pi(x+y) \geq \min(\mathrm{ord}_\pi(x), \mathrm{ord}_\pi(y))$.
c) Equality holds in part b) if $\mathrm{ord}_\pi(x) \neq \mathrm{ord}_\pi(y)$.

PROOF. We will suppose for simplicity that $x, y \in R \setminus \{0\}$. The general case follows by clearing denominators as usual. Put $a = \mathrm{ord}_\pi(x)$, $b = \mathrm{ord}_\pi(y)$. By hypothesis, there exists x', y' such that $x = \pi^a x'$, $y = \pi^b y'$ and $\pi \nmid x', y'$.
a) $xy = \pi^{a+b}(x'y')$. Thus $\mathrm{ord}_\pi(xy) \geq a + b$. Conversely, suppose that $\pi^{a+b+1} \mid xy$. Then $\pi \mid x'y'$, and, since π is a prime element, this implis $\pi \mid x'$ or $\pi \mid y'$, contradiction. Thus $\mathrm{ord}_\pi(xy) = a + b = \mathrm{ord}_\pi(x) + \mathrm{ord}_\pi(y)$.
b) Let $c = \min a, b$, so $x + y = \pi^c(\pi^{a-c}x' + \pi^{b-c}y'$, and thus $\pi^c \mid x + y$ and $\mathrm{ord}_\pi(x+y) \geq c = \min(\mathrm{ord}_\pi(x), \mathrm{ord}_\pi(y))$.
c) Suppose without loss of generality that $a < b$, and write $x + y = \pi^a(x' + \pi^{b-a}y')$. If $\pi^{a+1} \mid x + y = \pi^a x' + \pi^b y'$, then $\pi \mid x' + \pi^{b-a}y'$. Since $b - a > 0$, we have $\pi \mid (x' + \pi^{b-a}y') - (\pi^{b-a}y') = x'$, contradiction. □

Suppose that π and π' are associate nonzero prime elements, i.e., there exists a unit $u \in R$ such that $\pi' = u\pi$. Then a moment's thought shows that the ord functions ord_π and $\mathrm{ord}_{\pi'}$ coincide. This means that ord_π depends only on the principal ideal $\mathfrak{p} = (\pi)$ that the prime element π generates. We could therefore redefine the ord function as $\mathrm{ord}_\mathfrak{p}$ for a nonzero principal prime ideal $\mathfrak{p} = (\pi)$ of R, but for our purposes it is convenient to just choose one generator π of each such ideal $\mathfrak{p}$. Let $\mathcal{P}$ be a maximal set of mutually nonassociate nonzero prime elements, i.e., such that each nonzero prime ideal $\mathfrak{p}$ contains exactly one element of $\mathcal{P}$.

Now suppose that R is a UFD, and $x \in R \setminus \{0\}$ is an element such that $\mathrm{ord}_\pi(x) = 0$ for all $\pi \in \mathcal{P}$. Then x is not divisible by any irreducible elements, so is necessarily a unit. In fact the same holds for elements $x \in K \setminus \{0\}$, since we can express $x = \frac{a}{b}$ with a and b not both divisible by any prime element. (In other words, in a UFD we can reduce fractions to lowest terms!) It follows that any $x \in K \setminus \{0\}$ is determined

[1]We say that a properly divides b if $a \mid b$ but a is not associate to b.

up to a unit by the integers $\mathrm{ord}_\pi(x)$ as π ranges over elements of $\mathcal{P}$. Indeed, put

$$y = \prod_{\pi \in \mathcal{P}} \pi^{\mathrm{ord}_\pi x}.$$

Then we have $\mathrm{ord}_\pi(\frac{x}{y}) = 0$ for all $\pi \in \mathcal{P}$, so that $\frac{x}{y} = u$ is a unit in R, and $x = yu$.

After these preparations, the proof of Proposition 6.4 is straightforward: we have $xy = z^n$. For any prime element p, take ord_p of both sides to get

$$\mathrm{ord}_p(x) + \mathrm{ord}_p(y) = n \, \mathrm{ord}_p(z).$$

But since x and y are assumed coprime, for any fixed prime p, we have either $\mathrm{ord}_p(x) = 0$ or $\mathrm{ord}_p(y) = 0$. Either way we get that $n \mid \mathrm{ord}_p(x)$ and $n \mid \mathrm{ord}_p(y)$ (since $n \mid 0$ for all n). So the following are well-defined elements of R:

$$x' = \prod_{p \in \mathcal{P}} p^{\frac{\mathrm{ord}_p(x)}{n}},$$

$$y' = \prod_{p \in \mathcal{P}} p^{\frac{\mathrm{ord}_p(y)}{n}},$$

where the product extends over a maximal set of pairwise nonassociate nonzero prime elements of R. By construction, we have $\mathrm{ord}_p((x')^n) = n \, \mathrm{ord}_p(x') = n \frac{\mathrm{ord}_p(x)}{n} = \mathrm{ord}_p(x)$ for all $p \in \mathcal{P}$, so the elements x and $(x')^n$ are associate: i.e., there exists a unit u in R such that $x = u(x')^n$. Exactly the same applies to y and y': there exists a unit $v \in R$ such that $y = v(y')^n$.

3.2. Application to the Bachet-Fermat Equation.

To complete the proof of Theorem 6.3 we need to verify that the hypotheses of Proposition 6.4b) apply: namely, that every unit in $\mathbb{Z}[\sqrt{-2}]$ is a cube and that the elements $y + \sqrt{-2}$, $y - \sqrt{-2}$ are indeed relatively prime. For the former, we are fortunate in that, as for $\mathbb{Z}$, the only units in $R = \mathbb{Z}[\sqrt{-2}]$ are ± 1, both of which are indeed cubes in R.

For the latter, we suppose that $d \in R$ is a common divisor of $y + \sqrt{-2}$ and $y - \sqrt{-2}$. Then also $d \mid (y + \sqrt{-2}) - (y - \sqrt{-2}) = 2\sqrt{-2}$, i.e., there exists $d' \in R$ with $dd' = 2\sqrt{-2}$. Taking norms of both sides we get

$$N(d)N(d') = N(2\sqrt{-2}) = 8,$$

so $N(d) \mid 8$. Moreover, there exists $\alpha \in R$ such that $d\alpha = y + \sqrt{-2}$, hence

$$N(d)N(\alpha) = N(d\alpha) = N(y + \sqrt{2}) = y^2 + 2 = x^3,$$

so $N(d) \mid x^3$. We claim that x must be odd. For if not, then reducing the equation $x^3 = y^2 + 2 \bmod 8$ gives $y^2 \equiv 6 \pmod 8$, but the only squares mod 8 are $0, 1, 4$. Thus x^3 is odd and $N(d) \mid \gcd(x^3, 8) = 1$ so $d = \pm 1$ is a unit in R.

3.3. Application to the Mordell Equation with $k = 1$.

THEOREM 6.6. *The only integer solution to $y^2 + 1 = x^3$ is $(1, 0)$.*

PROOF. This time we factor the left hand side over the UFD $R = \mathbb{Z}[\sqrt{-1}]$:

$$(y + \sqrt{-1})(y - \sqrt{-1}) = x^3.$$

If a nonunit d in R divides both $y+\sqrt{-1}$ and $y-\sqrt{-1}$, then it divides $(y+\sqrt{-1})-(y-\sqrt{-1}) = 2\sqrt{-1} = (1+\sqrt{-1})^2\sqrt{-1}$. The element $1+\sqrt{-1}$, having norm $N(1+\sqrt{-1}) = 2$ a prime number, must be an irreducible (hence prime) element of R. So $1+i$ is the only possible common prime divisor. We compute

$$\frac{y \pm \sqrt{-1}}{1+\sqrt{-1}} \cdot \frac{1-\sqrt{-1}}{1-\sqrt{-1}} = \frac{y \pm 1 + (y \pm 1)\sqrt{-1}}{2},$$

which is an element of R iff y is odd. But consider the equation $y^2+1 = x^3$ modulo 4: if y is odd, then $y^2+1 \equiv 2 \pmod 4$, but 2 is not a cube modulo 4. Therefore we must have that y is even, so that $y \pm \sqrt{-1}$ are indeed coprime. Moreover, although the unit group of R is slightly larger in this case – it is $\{\pm 1, \pm\sqrt{-1}\}$ – it is easily checked that every unit is a cube in R. So Proposition 6.4b) applies here, giving $\alpha, \beta \in R$ such that

$$y + \sqrt{-1} = \alpha^3, \ y - \sqrt{-1} = \beta^3.$$

Again we will put $\alpha = a + b\sqrt{-1}$ and expand out α^3, getting

$$y + \sqrt{-1} = a^3 - 3b^2a + (3a^2b - b^3)\sqrt{-1},$$

or

$$y = a(a^2 - 3b^2), \ 1 = b(3a^2 - b^2).$$

So we have either $1 = b = 3a^2 - b^2$, which leads to $3a^2 = 2$, which has no integral solution, or $-1 = b = 3a^2 - b^2$, which leads to $a = 0$, so $\alpha = -\sqrt{-1}$, $y = (-\sqrt{-1})^3 - \sqrt{-1} = 0$, $x = 1$ and thus to $(x, y) = (1, 0)$. □

4. Beyond UFDs

The situation here is somewhat analogous to our study of the equations $x^2+Dy = p$, where the assumption that he quadratic ring $\mathbb{Z}[\sqrt{-D}]$ is a UFD leads to a complete solution of the problem. However there are also some differences. First, whereas in the present situation we are using the assumption that $\mathbb{Z}[\sqrt{-k}]$ is a UFD in order to show that $y^2 + k = x^3$ has very few solutions, earlier we used the assumption that $\mathbb{Z}[\sqrt{D}]$ is a UFD to show that the family of equations $x^2 + Dy^2 = p$ had many solutions, namely for all primes p for which $-D$ is a square mod p.

A more significant difference is that the assumption $\mathbb{Z}[\sqrt{D}]$ was necessary as well as sufficient for our argument to go through: we saw that whenever $D < -3$ 2 is not of the form x^2+Dy^2. On the other hand, suppose $\mathbb{Z}[\sqrt{-k}]$ is not a UFD: must the coprime powers trick fail? It is not obvious, so let us study it more carefully.

We would like to axiomatize the coprime powers trick. There is an agreed upon definition of coprimality of two elements x and y in a general domain R: if $d \mid x$ and $d \mid y$ then d is a unit. However it turns out to be convenient to require a stronger property than this, namely that the ideal $\langle x, y\rangle = \{rx + sy \mid r, s \in R\}$ generated by x and y be the unit ideal R. More generally, for two ideals I, J of a ring, the sum $I + J = \{i + j \mid i \in I, j \in J\}$ is an ideal, and we say that I and J are **comaximal** if $I + J = R$; equivalently, the only ideal which contains both I and J is the "improper" ideal R. Since every proper ideal in a ring is contained in a maximal, hence prime, ideal, the comaximality can be further reexpressed as the property that there is no prime ideal $\mathfrak{p}$ containing both I and J. (This will be the formulation which is most convenient for our application.)

Notice that the condition that x and y be coprime can be rephrased as saying

that the only *principal* ideal (d) containing both x and y is the improper ideal $R = (1)$. So the notions of coprime and comaximal elements coincide in a principal domain, but not in general.

Now, for a positive positive integer n, say that an integral domain R has property **CM(n)** if the comaximal powers trick is valid in degree n: namely, for all $x, y, z \in R$ with $\langle x, y\rangle = R$ and $xy = z^n$, then there exist elements $a, b \in R$ and units $u, v \in R$ such that $x = ua^n$, $y = vb^n$. Exactly as above, if we also have $(R^\times)^n = (R^\times)$ – i.e., every unit in R is an nth power – then the units u and v can be omitted. Now consider the following

THEOREM 6.7. *Let $k \in \mathbb{Z}^+$ be squarefree with $k \equiv 1, 2 \pmod 4$. Suppose that the ring $\mathbb{Z}[\sqrt{-k}]$ has property CM(3). Then:*
a) If there exists an integer a such that $k = 3a^2 \pm 1$, then the only integer solutions to the Mordell equation $y^2 + k = x^3$ are $(a^2 + k, \pm a(a^2 - 3k))$.
b) If there is no integer a as in part a), the Mordell equation $y^2 + k = x^3$ has no integral solutions.

PROOF. [**IR**, Prop. 17.10.2] Suppose (x, y) is an integral solution to $y^2+k = x^3$. Reduction mod 4 shows that x is odd. Also $\gcd(k, x) = 1$: otherwise there exists a prime p dividing both k and x, so $p \mid x^3 - k = y^2$ and $p \mid y^2 \implies p^2 \mid x^3 - y^2 = k$, contradicting the squarefreeness of k. Now consider

$$(y + \sqrt{-k})(y - \sqrt{-k}) = x^3.$$

We wish to show that $\langle y + \sqrt{-k}, y - \sqrt{-k}\rangle = R$. If not, there exists a prime ideal $\mathfrak{p}$ of R with $y \pm \sqrt{-k} \in \mathfrak{p}$. Then $(y + \sqrt{-k}) - (y - \sqrt{-k}) = 2\sqrt{-k} \in \mathfrak{p}$, hence also $-(2\sqrt{-k})^2 = 4k \in \mathfrak{p}$. Moreover $\mathfrak{p}$ contains $y^2 + k = x^3$ and since it is prime, it contains x. But since x is odd and $\gcd(x, k) = 1$, also $\gcd(x, 4k) = 1$, so that there exist $m, n \in \mathbb{Z}$ with $1 = xm + 4kn$ and thus $1 \in \mathfrak{p}$. Moreover, either $k = 1$ (a case which we have already treated) or $k > 1$ and the only units of $\mathbb{Z}[\sqrt{-k}]$ are ± 1. Therefore there exists $\alpha = a + b\sqrt{-k} \in R$ such that

$$y + \sqrt{-k} = \alpha^3 = (a + b\sqrt{-k})^3 = a(a^2 - 3kb^2) + b(3a^2 - kb^2)\sqrt{-k}.$$

So $b = \pm 1$ and $k = db^2 = 3a^2 \pm 1$. The integer a determined by this equation is unique up to sign. So $y = \pm a(a^2 - 3k)$, and one easily computes $x = a^2 + k$. □

Since property CM(3) holds in the PIDs $\mathbb{Z}[\sqrt{-1}]$ and $\mathbb{Z}[\sqrt{-2}]$, whatever else Theorem 6.7 may be good for, it immediately implies Theorems 6.3 and 6.6. Moreover its proof was shorter than the proofs of either of these theorems! The economy was gained by consideration of not necessarily principal ideals.

Thus, if for a given k as in the statement of Theorem 6.7 we can find more solutions to the Mordell Equation than the ones enumerated in the conclusion of the theorem we know that $\mathbb{Z}[\sqrt{-k}]$ does not satisfy property CM(3). In the following examples we simply made a brute force search over all x and y with $|x| \leq 10^6$. (There is, of course, no guarantee that we will find *all* solutions this way!)

Example: The equation $y^2 + 26 = x^3$ has solutions $(x, y) = (3, \pm 1),\ (35, \pm 207)$, so $\mathbb{Z}[\sqrt{-26}]$ does not have property CM(3).

Example: The equation $y^2 + 53 = x^3$ has solutions $(x, y) = (9, \pm 26),\ (29, \pm 156)$, so $\mathbb{Z}[\sqrt{-53}]$ does not have CM(3).

Example: The equation $y^2 + 109 = x^3$ has solutions $(x, y) = (5, \pm 4),\ (145, \pm 1746)$. It is not trivial to find the solution $(145, \pm 1746)$ by hand, so perhaps it is easier to observe that 5 is not of the form $a^2 + 109$, $\mathbb{Z}[\sqrt{-109}]$, so by Theorem 6.7, $\mathbb{Z}[\sqrt{-109}]$ does not have property CM(3).

In fact, whether a ring $\mathbb{Z}[\sqrt{-k}]$ (here we keep the assumptions on k of Theorem 6.7, so in particular $\mathbb{Z}[\sqrt{-k}]$ is the full ring of algebraic integers of the quadratic field $\mathbb{Q}(\sqrt{-k})$; this would not be the case if $k \equiv 3 \pmod 4$) has property CM(k) can be determined algorithmically. It depends on an all-important numerical invariant called the **class number** of $\mathbb{Z}[\sqrt{-k}]$.

For any integral domain R, we can define an equivalence relation on the nonzero ideals of R. Namely, we decree that $I \sim J$ iff there exist $a, b \in R \setminus \{0\}$ such that $(a)I = (b)J$. Roughly speaking, we regard two ideals as being principal if and only if they differ multiplicatively from a principal ideal. When there are only finitely many equivalence classes, we define the **class number** of R to be the number of equivalence classes.[2] For example, if every ideal of R is principal, then the class number is equal to 1. Conversely, if the class number of R is equal to 1 and I is any nonzero ideal of R, then there exist a, b such that $aI = bR$. Then $b = b \cdot 1 \in aI$, so for some $x \in I$, $ax = b$. In particular $a \mid b$, and it is then easy to see that $I = (\frac{b}{a})R$. Thus the domains with class number one are precisely the principal ideal domains.

Now let K be a number field, and let $\mathbb{Z}_K$ be the ring of integers in K. In particular this includes $\mathbb{Z}[\sqrt{-k}]$ for k as above.

THEOREM 6.8. *Let $\mathbb{Z}_K$ be the ring of integers in a number field K. Then:*
a) There are only finitely many equivalence classes of ideals of $\mathbb{Z}_K$, so there is a well-defined class number, denoted $h(K)$.
b) The ring $\mathbb{Z}_K$ is a PID iff it is a UFD iff $h(K) = 1$.
c) Let $n \in \mathbb{Z}^+$. If $\gcd(n, h(K)) = 1$, then $\mathbb{Z}_K$ has property CM(n).

At several points in this course we have flirted with crossing the border into the land of algebraic number theory, but that no such passport is required is one of our ground rules. Because of this it is simply not possible to prove Theorem 6.8 here. We can only say that the study of such properties of the ring $\mathbb{Z}_K$ is a central topic in the classical theory of algebraic numbers.

Moreover, algorithms for computing the class number have been a very active part of algebraic number theory for more than one hundred years. Such algorithms are available – indeed, they have been implemented in many software packages – the question is only of the speed and memory needed to do the job. The case of (imaginary) quadratic fields is especially classical and relates to (positive definite) binary quadratic forms. So the following table of class numbers of $\mathbb{Q}(\sqrt{-k})$ for squarefree

[2]As we have stated it, the definition makes sense for arbitrary domains and is equivalent to the usual definition for number rings $\mathbb{Z}_K$. For more general domains – and even some quadratic rings – there is another (less elementary) definition which is more useful.

k, $1 \leq k \leq 200$ is more than two hundred years old:

$h(\mathbb{Q}(\sqrt{-k})) =$

1 for $k = 1, 2, 3, 7, 11, 19, 43, 67, 163$
2 for $k = 5, 6, 10, 13, 15, 22, 35, 37, 51, 58, 91, 115, 123, 187$
3 for $k = 23, 31, 59, 83, 107, 139$
4 for $k = 14, 17, 21, 30, 33, 34, 39, 42, 46, 55, 57, 70, 73, 78, 82, 85, 93, 97, 102, 130,$
$133, 142, 155, 177, 190, 193, 195$
5 for $k = 47, 79, 103, 127, 131, 179$
6 for $k = 26, 29, 38, 53, 61, 87, 106, 109, 118, 157$
7 for $k = 71, 151$
8 for $k = 41, 62, 65, 66, 69, 77, 94, 98, 105, 113, 114, 137, 138, 141, 145, 154, 158, 165, 178$
9 for $k = 199$
10 for $k = 74, 86, 122, 166, 181, 197$ 11 for $k = 167$
12 for $k = 89, 110, 129, 170, 174, 182, 186$
13 for $k = 191$
14 for $k = 101, 134, 149, 173$
16 for $k = 146, 161, 185$
20 for $k = 194$

So Theorem 6.7 applies to give a complete solution to the Mordell equation $y^2 + k = x^3$ for the following values of k:

$$1, 2, 5, 6, 10, 13, 14, 17, 21, 22, 30, 33, 34, 37, 41, 42, 46, 57, 58, 62, 65, 69, 70, 73, 74, 77, 78,$$
$$82, 85, 86, 93, 94, 97, 98, 101, 102, 106, 113, 114, 122, 130, 133, 134, 137, 138,$$
$$141, 142, 145, 146, 149, 154, 158, 161, 165, 166, 177, 178, 181, 185, 190, 193, 194, 197.$$

Example: The equation $y^2+47 = x^3$ has solutions $(x, y) = (6, \pm 13),\ (12, \pm 41), (63, \pm 500)$. On the other hand $\mathbb{Z}[\sqrt{-47}]$ has class number 5 so does have property CM(3). Note that $47 \equiv 3 \pmod 4$.

Example: $\mathbb{Z}[\sqrt{-29}]$ has class number 6, but nevertheless $y^2 + 29 = x^3$ has no integral solutions.[3] Thus there is (much) more to this story than the coprime powers trick. For more details, we can do no better than recommend [**M**, Ch. 26].

5. Remarks and Acknowledgements

Our first inspiration for this material was the expository note [**Conr-A**]. Conrad proves Theorems 6.2 and 6.3 as an application of unique factorization in $\mathbb{Z}$ and $\mathbb{Z}[\sqrt{-2}]$. Many more examples of successful (and one unsuccessful!) solution of Mordell's equation for various values of k are given in [**Conr-B**]. A range of techniques is showcased, including the coprime powers trick but also: elementary (but somewhat intricate) congruence arguments and quadratic reciprocity.

Also useful for us were lecture notes of P. Stevenhagen [**St-ANT**]. Stevenhagen's treatment is analogous our discussion of quadratic rings. In particular, he first

[3]How do we know? For instance, we can look it up on the internet:
`http://www.research.att.com/~njas/sequences/A054504`

proves Theorem 6.6. He then assumes that $\mathbb{Z}[\sqrt{-19}]$ satisfies CM(3) and deduces that $y^2+19 = x^3$ has no integral solutions; finally he points out $(x, y) = (18, 7)$. We did not discuss this example in the text because it depends critically on the fact that $\mathbb{Z}[\sqrt{-19}]$ is not the full ring of integers in $K = \mathbb{Q}(\sqrt{-19})$: rather $\mathbb{Z}_K = \mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$. For rings like $\mathbb{Z}[\sqrt{-19}]$ the definition we gave of the class number is not the correct one: we should count only equivalence classes of **invertible ideals** – i.e., nonzero ideals I for which there exists J such that IJ is principal. In this amended sense the class number of $\mathbb{Z}[\sqrt{-19}]$ is 3.

A generalization of Theorem 6.7 appears in §5.3 of lecture notes of Franz Lemmermeyer:

`http://www.fen.bilkent.edu.tr/~franz/ant/ant1-7.pdf`

Lemmermeyer finds all integer solutions to the equation $y^2 + k = x^3$ whenever $3 \nmid h(\mathbb{Q}(\sqrt{-k}))$ and $k \not\equiv 7 \pmod 8$. Again we have avoided this case so as not to have to deal with the case where $\mathbb{Z}[\sqrt{-k}]$is not the full ring of integers.

It is interesting to look at the work which has been done on the Mordell equation since Mordell's death in 1972. In 1973, London and Finkelstein [**LF73**] found all solutions to Mordell's equation for $|k| \leq 10^2$. The current state of the art is another story entirely: a 1998 work of Gebel, Pethö and Zimmer [**GPZ98**] solves the Mordell equation for $|k| \leq 10^4$ and for about 90% of integers k with $|k| \leq 10^5$.

CHAPTER 7

The Pell Equation

1. Introduction

Let D be a nonzero integer. We wish to find all integer solutions (x, y) to

$$x^2 - Dy^2 = 1. \tag{32}$$

1.1. History.

Leonhard Euler called (32) **Pell's Equation** after the English mathematician John Pell (1611-1685). This terminology has persisted to the present day, despite the fact that it is well known to be mistaken: Pell's only contribution to the subject was the publication of some partial results of Wallis and Brouncker. In fact the correct names are the usual ones: the problem of solving the equation was first considered by Fermat, and a complete solution was given by Lagrange.

By any name, the equation is an important one for several reasons – only some of which will be touched upon here – and its solution furnishes an ideal introduction to an entire branch of number theory, **Diophantine Approximation**.

1.2. First remarks on Pell's equation.

We call a solution (x, y) to (32) **trivial** if $xy = 0$. We always have at least two trivial solutions: $(x, y) = (\pm 1, 0)$, which we shall call **trivial**. As for any plane conic curve, as soon as there is one solution there are infinitely many *rational solutions* $(x, y) \in \mathbb{Q}^2$, and all arise as follows: draw all lines through a single point, say $(-1, 0)$, with rational slope r, and calculate the second intersection point (x_r, y_r) of this line with the quadratic equation (32).

The above procedure generates all rational solutions and thus contains all integer solutions, but figuring out which of the rational solutions are integral is not straightforward. This is a case where the question of integral solutions is essentially different, and more interesting, than the question of rational solutions. Henceforth when we speak of 'solutions" (x, y) to (32) we shall mean integral solutions.

Let us quickly dispose of some uninteresting cases.

PROPOSITION 7.1. *If the Pell equation $x^2 - Dy^2 = 1$ has nontrivial solutions, then D is positive and not a perfect square.*

PROOF. • $(D = -1)$: The equation $x^2 + y^2 = 1$ has four trivial solutions: $(\pm 1, 0), (0, \pm 1)$.
• $(D < -1)$: Then $x \neq 0 \implies x^2 - dy^2 \geq 2$, so (32) has only the solutions $(\pm 1, 0)$.
• $(D = N^2)$: Then $x^2 - Dy^2 = (x + Ny)(x - Ny) = 1$, and this necessitates either:

$$x + Ny = x - Ny = 1$$

in which case $x = 1$, $y = 0$; or

$$x + Ny = x - Ny = -1,$$

in which case $x = -1, y = 0$: there are only trivial solutions. □

From now on we assume that D is positive and not a perfect square.

Now observe that nontrivial solutions come in quadruples: if (x, y) is any one solution, so is $(-x, y)$, $(x, -y)$ and $(-x, -y)$. We describe these solutions by a pair of signs: $(+, +)$, $(+, -)$, $(-, +)$, $(-, -)$.

2. Example: The equation $x^2 - 2y^2 = 1$

Let us take $D = 2$. The equation $x^2 - 2y^2 = 1$ can be rewritten as

$$y^2 = \frac{x^2 - 1}{2}.$$

In other words, we are looking for positive integers x for which $\frac{x^2-1}{2}$ is an integer square. First of all $x^2 - 1$ must be even, so x must be odd. Trying $x = 1$ gives, of course, the trivial solution $(1, 0)$. Trying $x = 3$ we get

$$\frac{3^2 - 1}{2} = 4 = 2^2,$$

so $(3, 2)$ is a $(+, +)$ solution. Trying successively $x = 5, 7, 9$ and so forth we find that it is rare for $\frac{x^2-1}{2}$ to be a square: the first few values are 12, 24, 40, 60, 69, 112 and then finally with $x = 17$ we are in luck:

$$\frac{17^2 - 1}{2} = 144 = 12^2,$$

so $(17, 12)$ is another $(+, +)$ solution. Searching for further solutions is a task more suitable for a computer. My laptop has no trouble finding some more solutions: the next few are $(99, 70)$, $(577, 408)$, and $(3363, 2378)$. Further study suggests that (i) the equation $x^2 - 2y^2$ has infinitely many integral solutions, and (ii) the size of the solutions is growing rapidly, perhaps even exponentially.

2.1. Return of abstract algebra. Hopefully we have not forgotten the connection between $x^2 - Dy^2$ and the quadratic ring $\mathbb{Z}[\sqrt{D}]$. Namely, we have the conjugation operation

$$\alpha = x + y\sqrt{D} \mapsto \overline{\alpha} = x - y\sqrt{D}$$

and

$$N(\alpha) = N(x + y\sqrt{D}) = (x + y\sqrt{D})(x - y\sqrt{D}) = x^2 - Dy^2.$$

Thus the solutions to the Pell equation $x^2 - Dy^2 = 1$ are the precisely the elements of norm 1 in the quadratic ring. These are all units of the ring $\mathbb{Z}[\sqrt{D}]$. (By taking $D = 2$ and $D = 3$, we saw that there may or may not also be units of norm -1. Thus the "negative Pell equation" $x^2 - Dy^2 = -1$ is also interesting....in fact it turns out to be more interesting and difficult than the Pell equation itself. We will make a few remarks about it at the end of the chapter.) When $D < 0$ it was an easy exercise to find all (finitely many) units in $\mathbb{Z}[\sqrt{D}]$. The case of $D > 0$ is considerably more interesting!

2.2. The solution for $D = 2$.

In any ring R, the units $R^\times$ form a group under multiplication. Moreover, by the multiplicativity of the norm map, the units of norm one form a subgroup of the entire unit group $\mathbb{Z}[\sqrt{D}]^\times$.[1] This is really a key observation, because it allows us to *multiply* solutions to the Pell equation: if $x_1^2 - Dy_1^2 = 1$ and $x_2^2 - Dy_2^2 = 1$, then

$$1 = 1 \cdot 1 = N(x_1 + y_1\sqrt{D})N(x_2 + y_2\sqrt{D}) = N((x_1 + y_1\sqrt{D})(x_2 + y_2\sqrt{D}))$$

$$= N(x_1x_2 - Dy_1y_2 + (x_1y_2 + x_2y_1)\sqrt{D}) = (x_1x_2 - Dy_1y_2)^2 - D(x_1y_2 + x_2y_1)^2.$$

Let us try out this formula in the case $D = 2$ for $(x_1, y_1) = (x_2, y_2) = (3, 2)$. Our new solution is $(3 \cdot 3 + 2 \cdot 2 \cdot 2, 2 \cdot 3 + 3 \cdot 2) = (17, 12)$, nothing else than the second smallest positive solution! If we now apply the formula with $(x_1, y_1) = (17, 12)$ and $(x_2, y_2) = (3, 2)$, we get the next smallest solution $(99, 70)$.

Indeed, for any positive integer n, we may write the nth power $(3 + 2\sqrt{2})^n$ as $x_n + y_n\sqrt{2}$ and know that (x_n, y_n) is a solution to the Pell equation. One can see from the formula for the product that it is a positive solution. Moreover, the solutions are all different because the real numbers $(3 + 2\sqrt{2})^n$ are all distinct. We get the trivial solution $(1, 0)$ by taking the 0th power of $3 + 2\sqrt{2}$. Moreover, $(3 + 2\sqrt{2})^{-1} = 3 - 2\sqrt{2}$ is a "half-positive" solution, and taking negative integral powers of $3 + 2\sqrt{2}$ we get infinitely many more such solutions.

In total, every solution to $x^2 - 2y^2 = 1$ that we have found is of the form $\pm(x_n, y_n)$ where $x_n + y_n\sqrt{2} = (3 + 2\sqrt{2})^n$ for some $n \in \mathbb{Z}$.

Let us try to prove that these are *all* the integral solutions. It is enough to show that every $(+, +)$ solution is of the form (x_n, y_n) for some positive integer n, since every norm one element $x + y\sqrt{d}$ is obtained from an element with $x, y \in \mathbb{Z}^+$ by multiplying by -1 and/or taking the reciprocal.

LEMMA 7.2. *Let (x, y) be a nontrivial integral solution to $x^2 - Dy^2 = 1$. Then:*
a) We have $x > 0$ and $y > 0 \iff x + y\sqrt{d} > 1$.
b) We have $x > 0$ and $y < 0 \iff 0 < x + y\sqrt{d} < 1$.
c) We have $x < 0$ and $y > 0 \iff -1 < x + y\sqrt{d} < 0$.
d) We have x and y are both negative $\iff x + y\sqrt{d} < -1$.

PROOF. Exercise. □

For any real $M > 1$, we observe that there can only be finitely many pairs $(x, y) \in \mathbb{Z}^+$ such that $x + y\sqrt{D} \leq M$. Indeed, if $x + y\sqrt{D} \leq M$ then we must have $1 \leq x, y \leq M$, so there are at most M^2 possibilities. We have $3 + 2\sqrt{2} \leq 6$; checking the 36 integers $x, y \in [1, 6]$ we find that $3 + 2\sqrt{2}$ is the smallest $(+, +)$ solution.

Let $x, y \in \mathbb{Z}^+$ be such that $x^2 - 2y^2 = 1$. There is a largest $n \in \mathbb{N}$ such that

$$x + y\sqrt{2} \geq (3 + 2\sqrt{2})^n.$$

Put

$$\alpha := (x + y\sqrt{2})(3 + 2\sqrt{2})^{-n}.$$

[1] It is easy to see that the norm one subgroup has index 2 if there exists a unit of norm -1 and is equal to the unit group otherwise.

Then

$$N(\alpha) = N(x + y\sqrt{2})N(3 + 2\sqrt{2})^{-n} = 1.$$

By our choice of n, we have $\alpha \geq 1$. But moreover we have $\alpha < 3 + 2\sqrt{2}$, since if $\alpha \geq 3 + 2\sqrt{2}$ then $\frac{\alpha}{3+2\sqrt{2}} \geq 1$ and thus $(x + y\sqrt{2}) \geq (3 + 2\sqrt{2}^{n+1}$, contradicting our definition of n. Put $\alpha = x' + y'\sqrt{D}$ If $\alpha > 1$, then (x', y') is a $(+, +)$ solution to the Pell equation with $x' + y'\sqrt{2} < 3 + 2\sqrt{2}$, contradicting what we showed above. So it must be that $\alpha = 1$ and thus

$$(x + y\sqrt{2} = (3 + 2\sqrt{2})^n.$$

This completes the proof.

Thus we have "solved the Pell equation" for $D = 2$. To add icing, we can give explicit formulas for the solutions. Namely, we know that every $(+, +)$ solution (x, y) is of the form

$$x_n + y_n\sqrt{2} = (3 + 2\sqrt{2})^n$$

for $n \in \mathbb{Z}^+$. If we apply conjugation to this equation, then using the fact that it is a field homomorphism, we get

$$x_n - y_n\sqrt{2} = (3 - 2\sqrt{2})^n.$$

dding the two equations and dividing by 2, we get

$$x_n = \frac{(3 + 2\sqrt{2})^n + (3 - 2\sqrt{2})^n}{2},$$

and similarly we solve for y_n, getting

$$y_n = \frac{(3 + 2\sqrt{2})^n + (3 - 2\sqrt{2}^n}{2\sqrt{2}}.$$

But wait, thre's more! Since $3 - 2\sqrt{2} = 0.17157\ldots$, for all $n \in \mathbb{Z}^+$ the terms $\frac{(3-2\sqrt{2})^n}{2}$ and $\frac{(3-2\sqrt{2})^n}{2\sqrt{2}}$ in x_n and y_n are exponentially decaying to 0 and are always less than $\frac{1}{2}$. For a real number α such that $\alpha - \frac{1}{2} \notin \mathbb{Z}$, there is a unique nearest integer to α that we denote by $\lfloor \alpha \rceil$. By what we said above, we can neglect the exponentially small terms simply by rounding to the nearest integer, getting the following result.

THEOREM 7.3. *Every solution to $x^2 - 2y^2 = 1$ with $x, y \in \mathbb{Z}^+$ is of the form*

$$x_n = \lfloor \frac{(3 + 2\sqrt{2})^n}{2} \rceil,$$

$$y_n = \lfloor \frac{(3 + 2\sqrt{2})^n}{2\sqrt{2}} \rceil$$

for some $n \in \mathbb{Z}^+$.

Among other things, this explains why it was not so easy to find solutions by hand: the size of both the x and y coordinates grow exponentially! The reader is invited to plug in a value of n for herself: for e.g. $n = 17$ it is remarkable how close the irrational numbers $u^{17}/2$ and $u^{17}/(2\sqrt{2})$ are to integers:

$$u^{17}/2 = 5168247530882.999999999999949;$$

$$u^{17}/(2\sqrt{2}) = 3654502875938.000000000000032.$$

A bit of reflection reveals that this has a lot to do with the fact that $\frac{x_n}{y_n}$ is necessarily very close to $\sqrt{2}$. Indeed, by turning this observation on its head we shall solve the Pell equation for general nonsquare d.

3. A result of Dirichlet

LEMMA 7.4. *(Dirichlet) For any irrational (real) number α, there are infinitely many rational numbers $\frac{x}{y}$ (with $\gcd(x,y)=1$) such that*

$$|\alpha - \frac{x}{y}| < \frac{1}{y^2}.$$

PROOF. Since the lowest-term denominator of any rational number $\frac{x}{y}$ is unchanged by subtracting any integer n, by subtracting the integer part $[\alpha]$ of α we may assume $\alpha \in [0,1)$. Now divide the half-open interval $[0,1)$ into n equal pieces: $[0,\frac{1}{n}) \cup [\frac{1}{n},\frac{2}{n}) \ldots \cup [\frac{n-1}{n},1)$. Consider the fractional parts of $0, \alpha, 2\alpha, \ldots, n\alpha$. Since we have $n+1$ numbers in $[0,1)$ and only n subintervals, by the pigeonhole principle some two of them must lie in the same subinterval. That is, there exist $0 \leq j < k \leq n$ such that

$$|k\alpha - [k\alpha] - (j\alpha - [j\alpha])| < \frac{1}{n}.$$

Now take $y = k - j$, $x = [k\alpha] - [j\alpha]$, so that the previous inequality becomes

$$|y\alpha - x| < \frac{1}{n}.$$

We may assume that $\gcd(x,y) = 1$, since were there a common factor, we could divide through by it and that would only improve the inequality. Moreover, since $0 < y \leq n$, we have

$$|\alpha - \frac{x}{y}| < \frac{1}{ny} < \frac{1}{y^2}.$$

This exhibits one solution. To see that there are infinitely many, observe that since α is irrational, the quantity $|\alpha - \frac{x}{y}|$ is always strictly greater than 0. But by choosing n sufficiently large we can apply the argument to find a rational number $\frac{x'}{y'}$ such that

$$|\alpha - \frac{x'}{y'}| < |\alpha - \frac{x}{y}|,$$

and hence there are infinitely many. □

Remark: The preceding argument is perhaps the single most famous application of the pigeonhole principle. Indeed, in certain circles, the pigeonhole principle goes by the name "Dirichlet's box principle"[2] because of its use in this argument.

4. Existence of Nontrivial Solutions

We are now ready to prove that for all positive nonsquare integers D, the Pell equation $x^2 - Dy^2 = 1$ has a nontrivial solution. Well, almost. First we prove an "approximation" to this result and then use it to prove the result itself.

PROPOSITION 7.5. *For some real number M, there exist infinitely many pairs of coprime positive integers (x,y) such that $|x^2 - Dy^2| < M$.*

[2]And in other circles, by the name "Schubfachprinzip."

PROOF. Applying Lemma 7.4 to $\alpha = \sqrt{D}$, we get an infinite sequence of coprime positive (since $\sqrt{D}$ is positive) integers (x, y) with $|\frac{x}{y} - \sqrt{D}| < \frac{1}{y^2}$. Multiplying through by y, the inequality is equivalent to

$$|x - y\sqrt{D}| < \frac{1}{y}.$$

Since

$$|x^2 - Dy^2| = |x - y\sqrt{D}||x + y\sqrt{D}|,$$

in order to bound the left hand side we also need a bound on $|x + y\sqrt{D}|$. There is no reason to expect that it is especially small, but using the triangle inequality we can get the following:

$$|x + \sqrt{D}y| = |x - \sqrt{D}y + 2\sqrt{D}y| \leq |x - \sqrt{D}y| + 2\sqrt{D}y < \frac{1}{y} + 2\sqrt{D}y.$$

Thus

$$|x^2 - Dy^2| < (\frac{1}{y})(\frac{1}{y} + 2\sqrt{D}y) = \frac{1}{y^2} + 2\sqrt{D} \leq 1 + 2\sqrt{D} = M.$$

□

THEOREM 7.6. *For any positive nonsquare integer D, the equation $x^2 - Dy^2 = 1$ has a nontrivial integral solution (x, y).*

PROOF. We begin by further exploiting the pigeonhole principle. Namely, since we have infinitely many solutions (x, y) to $|x^2 - Dy^2| < M$, there must exist some integer m, $|m| < M$ for which we have infinitely many solutions to the equality $x^2 - Dy^2 = m$. Observe that we cannot have $m = 0$, since $x^2 - Dy^2 = 0$ implies that $D = \frac{x^2}{y^2}$ is a perfect square. And now one more pigeonholing: we must have two different solutions, say (X_1, Y_1) and (X_2, Y_2) with $X_1 \equiv X_2 \pmod{|m|}$ and $Y_1 \equiv Y_2 \pmod{|m|}$ (since there are only m^2 different options altogether for $(x \pmod{|m|}, y \pmod{|m|})$ and infinitely many solutions). Let us write

$$\alpha = X_1 + Y_1\sqrt{S}$$

and

$$\beta = X_2 + Y_2\sqrt{D};$$

we have $N(\alpha) = N(\beta) = m$. A first thought is to divide α by β to get an element of norm 1; however, $\alpha/\beta \in \mathbb{Q}(\sqrt{D})$ but does not necessarily have integral x and y coordinates. However, it works after a small trick: consider instead

$$\alpha\overline{\beta} = X + Y\sqrt{D}.$$

I claim that both X and Y are divisible by m. Indeed we just calculate, keeping in mind that modulo m we can replace X_2 with X_1 and Y_2 with Y_1:

$$X = X_1X_2 - dY_1Y_2 \equiv X_1^2 - dY_1^2 \equiv 0 \pmod{|m|},$$

$$Y = X_1Y_2 - X_2Y_1 \equiv X_1Y_1 - X_1Y_1 \equiv 0 \pmod{|m|}.$$

Thus $\alpha\overline{\beta} = m(x + y\sqrt{D})$ with $x, y \in \mathbb{Z}$. Taking norms we get

$$m^2 = N(\alpha)N(\overline{\beta}) = N(\alpha\overline{\beta}) = N(m(x + y\sqrt{D})) = m^2(x^2 - Dy^2).$$

Since $m \neq 0$, this gives

$$x^2 - Dy^2 = 1.$$

Moreover $y \neq 0$: if $y = 0$ then the irrational part of Y, namely $X_1Y_2 - X_2Y_1$, would be zero, i.e., $\frac{X_1}{Y_1} = \frac{X_2}{Y_2}$, but this is impossible since $(X_1, Y_1) \neq (X_2, Y_2)$ are both coprime pairs: they cannot define the same rational number. We are done. □

5. The Main Theorem

Now we find **all solutions** of the Pell equation.

THEOREM 7.7. *Let D be a positive, nonsquare integer. Then:*
a) There are unique $x_1, y_1 \in \mathbb{Z}^+$ such that $x_1^2 - Dy_1^2 = 1$ and $x_1 + y_1\sqrt{D}$ is minimal. Put $u := x_1 + y_1\sqrt{D}$. Then every positive integral solution is of the form

$$(x_n, y_n) = \left(\frac{u^n + (u')^n}{2}, \frac{u^n - (u')^n}{2\sqrt{D}}\right) = \left(\lfloor \frac{u^n}{2} \rceil, \lfloor \frac{u^n}{2\sqrt{D}} \rceil\right)$$

for a unique $n \in \mathbb{Z}^+$.
b) Every solution to the Pell equation is of the form $\pm(x_n, y_n)$ for $n \in \mathbb{Z}$.

PROOF. Above we showed the existence of a nontrivial solution (x, y) and thus a $(+,+)$ solution. It is easy to see that for any $M > 0$ there are only finitely many pairs of positive integers such that $x + y\sqrt{D} \leq M$, so among all positive solutions, there must exist one with $x + y\sqrt{D}$ least. By taking positive integral powers of this fundamental solution $x_1 + y_1\sqrt{D}$ we get infinitely many positive solutions, whose x and y coordinates can be found explicitly as in §2. Moreover, the argument of §2 – given there for $D = 2$ – works generally to show that every positive solution is of this form. The reader is invited to look back over the details. □

6. A Caveat

It is time to admit that "solving the Pell equation" is generally taken to mean explicitly finding the fundamental solution $x_1 + y_1\sqrt{D}$. As usual in this course, we have concentrated on existence and not considered the question of how difficult it would be in practice to find the solution. Knowing that it exists we can, in principle, find it by trying all pairs (x, y) in order of increasing size. When $D = 2$ this is immediate. If we try other values of D we see that sometimes it is no trouble at all:

For $D = 3$, the fundamental solution is $(2, 1)$. For $D = 6$, it is $(5, 2)$. Similarly the fundamental solution can be found by hand for $D \leq 12$; it is no worse than $(19, 6)$ for $d = 10$. However, for $D = 13$ it is $(649, 180)$: a big jump!

If we continue to search we find that the size of the fundamental solution seems to obey no reasonable law: it does not grow in a steady way with D – e.g. for $D = 42$ it is the tiny $(13, 2)$ – but sometimes it is very large: for $D = 46$ it is $(24335, 3588)$, and – hold on to your hat! – for $D = 61$ the fundamental solution is

$$(1766319049, 226153980).$$

And things get worse from here on in: one cannot count on a brute-force search for D even of modest size (e.g. five digits).

There are known algorithms which find the fundamental solution relatively efficiently. The most famous and elementary of them is as follows: one can find the fundamental solution as a *convergent* in the continued fraction expansion of $\sqrt{D}$,

and this is relatively fast – it depends upon the *period length*. Alas, we shall not touch the theory of continued fractions in this course.

Continued fractions are not the last word on solving the Pell Equation, however. When D is truly large, other methods are required. Amazingly, a test case for this can be found in the mathematics of antiquity: the so-called **cattle problem of Archimedes**. Archimedes composed a lengthy poem ("twenty-two Greek elegiac distichs") which is in essence the hardest word problem in human history. The first part, upon careful study, reduces to solving a linear Diophantine equation (in several variables), which is essentially just linear algebra, and it turns out that there is a positive integer solution. However, to get this far is "merely competent", according to Archimedes. The second part of the problem poses a further constraint which boils down to solving a Pell equation with $D = 410286423278424$. In 1867 C.F. Meyer set out to solve the problem using continued fractions. However, he computed 240 steps of the continued fraction expansion of $\sqrt{410286423278424}$, whereas the period length is in fact 203254. Only in 1880 was the problem solved, by A. Amthor. (The gap between the problem and the solution – 2000 years and change – makes the case of Fermat's Last Theorem look fast!) Amthor used a different method. All of this and much more is discussed in a recent article by Hendrik Lenstra [**Le02**].

7. Some Further Comments

There is much more to be said on the subject. Just to further scratch the surface:

It is a purely algebraic consequence of our main result that the unit group of the ring $\mathbb{Z}[\sqrt{D}]$ (for D positive and nonsquare, as usual) is isomorphic to $\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Indeed, in solving the Pell equation, we found that the group of all norm one units is of this form, and it remains to account for units of norm -1. Sometimes there are none – e.g. when d is a prime which is $3 \pmod 4$ – and in this case the result is clear. But in any case the map $N : \mathbb{Z}[\sqrt{D}]^\times \to \{\pm 1\}$ has as its kernel the solutions to the Pell equation, so if there are also units of norm -1 the units of norm 1 form an index 2 subgroup. On the other hand units of finite order are necessarily roots of unity, of which there are no more than ± 1 in all of $\mathbb{R}$, let alone $\mathbb{Q}(\sqrt{D})$. The result follows from these considerations; the proof of this is left as an optional exercise.

This is a special case of an extremely important and general result in algebraic number theory. Namely, one can consider any *algebraic number field* – a finite degree field extension K of $\mathbb{Q}$ – and then the ring $\mathcal{O}_K$ of all algebraic integers of K – that is, elements α of K which satisfy a monic polynomial with $\mathbb{Z}$ coefficients. We have been looking at the case $K = \mathbb{Q}(\sqrt{d})$, a real quadratic field. Other relatively familiar examples are the cyclotomic fields $\mathbb{Q}(\zeta_N)$ obtained by adjoining an Nth root of unity: one can show that this field has degree $\varphi(N)$ over $\mathbb{Q}$ (equivalently, the cyclotomic polynomial Φ_N is irreducible over $\mathbb{Q}$).

Dirichlet's unit theorem asserts that the units $\mathcal{O}_K^\times$ form a finitely generated abelian group, i.e., are isomorphic to $\mathbb{Z}^a \times F$, where F is a finite group (which is in fact the group of roots of unity of F). Noting that the unit group is finite for imaginary quadratic fields and infinite for real quadratic fields, one sees that the rank a must

depend upon more than just the degree $d = [K : \mathbb{Q}]$ of the number field: somehow it depends upon "how real" the field is. More precisely, let r be the number of field homomorphisms from K into $\mathbb{R}$. Alternately, one can show that K is obtained by adjoining a single algebraic number α, i.e., $K = \mathbb{Q}[t]/(P(t))$, where P is a polynomial of degree $d = [K : \mathbb{Q}]$. Then r is nothing else than the number of real roots of the defining ("minimal") polynomial $P(t)$. In particular $r \leq d$, and $d - r$, the number of complex roots, is even. Then the precise form of Dirichlet's Unit Theorem asserts that $a = r + \frac{d-r}{2} - 1$, a quantity which is positive in every case except for $K = \mathbb{Q}$ and K an imaginary quadratic field! However the proof in the general case requires different techniques.

However, the argument that we used to find the general solution to the Pell equation is fascinating and important. On the face of it, it is very hard to believe that the problem of finding good rational approximations to an irrational number (a problem which is, let's face it, not initially so fascinating) can be used to solve Diophantine equations: we managed to use a result involving real numbers and inequalities to prove a result involving equalities and integers! This is nothing less than an entirely new tool, lying close to the border between algebraic and analytic number theory (and therefore helping to ensure a steady commerce between them). This subject is notoriously difficult – but here is one easy result. Define

$$\mathcal{L} = \sum_{n=0}^{\infty} 10^{-k!}.$$

We have a decimal expansion in which each lonely 1 is followed by a very long succession of zeros. The rational numbers afforded by the partial sums $\frac{p_N}{q_N} = \sum_{n=0}^{N} 10^{-k!}$ give excellent approximations: for any $A, B > 0$, one has

$$|\mathcal{L} - \frac{p_N}{q_N}| < \frac{A}{q_N^B}$$

for all sufficiently large N. On the other hand, Liouville proved the following:

THEOREM 7.8. *Suppose α satisfies a polynomial equation $a_d x^d + \ldots + a_1 x + a_0$ with $\mathbb{Z}$-coefficients. Then there is $A > 0$ such that for all integers p and $0 \neq q$,*

$$|\alpha - \frac{p}{q}| > \frac{A}{q^d}.$$

That is, being algebraic of degree d imposes an upper limit on the goodness of the approximation by rational numbers. An immediate and striking consequence is that Liouville's number $\mathcal{L}$ cannot satisfy an algebraic equation of any degree: that is, it is a transcendental number. In fact, by this argument Liouville established the existence of transcendental numbers for the first time!

Liouville's theorem was improved by many mathematicians, including Thue and Siegel, and culminating in the following theorem of Klaus Roth:

THEOREM 7.9. *(Roth, 1955) Let α be an algebraic real number, and let $\epsilon > 0$ be given. Then there are at most finitely many rational numbers $\frac{p}{q}$ satisfying*

$$|\alpha - \frac{p}{q}| < \frac{1}{q^{2+\epsilon}}.$$

For this result Roth won the Fields Medal in 1958.

CHAPTER 8

Arithmetic Functions

1. Introduction

Definition: An **arithmetic function** is a function $f : \mathbb{Z}^+ \to \mathbb{C}$.

Truth be told, this definition is a bit embarrassing. It would mean that taking any function from calculus whose domain contains $[1, +\infty)$ and restricting it to positive integer values, we get an arithmetic function. For instance, $\frac{e^{-3x}}{\cos^2 x + (17 \log(x+1))}$ is an arithmetic function according to this definition, although it is, at best, dubious whether this function holds any significance in number theory.

If we were honest, the definition we would like to make is that an arithmetic function is a real or complex-valued function defined for positive integer arguments which *is of some arithmetic significance*, but of course this is not a formal definition at all. Probably it is best to give examples:

EXAMPLE 8.1. *The prime counting function $n \mapsto \pi(n)$, the number of prime numbers p, $1 \leq p \leq n$.*

This is the example *par excellence* of an arithmetic function: approximately half of number theory is devoted to understanding its behavior. This function really deserves a whole unit all to itself, and it will get one: we put it aside for now and consider some other examples.

EXAMPLE 8.2. *The function $\omega(n)$, which counts the number of distinct prime divisors of n.*

EXAMPLE 8.3. *The function $\varphi(n)$, which counts the number of integers k, $1 \leq k \leq n$, with $\gcd(k, n) = 1$. Properly speaking this function is called **the totient function**, but its fame inevitably precedes it and modern times it is usually called just "the phi function" or "Euler's phi function." Since a congruence class $\overline{k}$ modulo n is invertible in the ring $\mathbb{Z}/n\mathbb{Z}$ iff its representative k is relatively prime to n, an equivalent definition is*

$$\varphi(n) := \#(\mathbb{Z}/n\mathbb{Z})^\times,$$

the cardinality of the unit group of the finite ring $\mathbb{Z}/n\mathbb{Z}$.

EXAMPLE 8.4. *The function $n \mapsto d(n)$, the number of positive divisors of n.*

EXAMPLE 8.5. *For any integer k, the function $\sigma_k(n)$, defined as*

$$\sigma_k(n) = \sum_{d \mid n} d^k,$$

the sum of the kth powers of the positive divisors of n. Note that $\sigma_0(n) = d(n)$.

EXAMPLE 8.6. *The Möbius function $\mu(n)$, defined as follows: $\mu(1) = 1$, $\mu(n) = 0$ if n is not squarefree; $\mu(p_1 \cdots p_r) = (-1)^r$, when $p_1, \ldots, p_r$ are distinct primes.*

EXAMPLE 8.7. *For a positive integer k, the function $r_k(n)$ which counts the number of representations of n as a sum of k integral squares:*

$$r_k(n) = \#\{(a_1, \ldots, a_k) \mid a_1^2 + \ldots + a_k^2 = n\}.$$

These examples already suggest many others. Notably, most of our examples are special cases of the following general construction: if we have on hand, for any positive integer n, a finite set S_n of arithmetic objects, then we can define an arithmetic function by defining $n \mapsto \#S_n$. This shows the link between number theory and combinatorics. In fact the Möbius function μ is a yet more purely combinatorial gadget, whose purpose we shall learn presently. In general we have lots of choices as to what sets S_n we want to count: the first few examples are "elementary" in the sense that the sets counted are defined directly in terms of such things as divisibility, primality, and coprimality: as we shall, see, they are also elementary in the sense that we can write down exact formulas for them. The example $r_k(n)$ is more fundamentally Diophantine in character: we have a polynomial in several variables – here $P(x_1, \ldots, x_k) = x_1^2 + \ldots + x_k^2$, and the sets we are conunting are just the number of times the value n is taken by this polynomial. This could clearly be much generalized, with the obvious proviso that there should be some suitable restrictions so as to make the number of solutions finite in number (e.g. we would not want to count the number of integer solutions to $ax + by = N$, for that is infinite; however we could restrict x and y to taking non-negative values). Ideally we would like to express these "Diophantine" arithmetic functions like r_k in terms of more elementary arithmetic functions like the divisor sum functions σ_k. Very roughly, this is the arithmetic analogue of the analytical problem expressing a real-valued function $f(x)$ as a combination of simple functions like x^k or $\cos(nx)$, $\sin(nx)$. Of course in analysis most interesting functions are not just polynomials (or trigonometric polynomials), at least not exactly: rather, one either needs to consider approximations to f by elementary functions, or to express f as some sort of limit (e.g. an infinite sum) of elementary functions (or both, of course). A similar philosophy applies here, with a notable exception: even the "elementary" functions like $d(n)$ and $\varphi(n)$ are not really so elementary as they first appear!

2. Multiplicative Functions

2.1. Definition and basic properties.

An important property shared by many "arithmetically significant" functions is multiplicativity.

Definition: An arithmetic function f is said to be **multiplicative** if:
(M1) $f(1) \neq 0$.
(M2) For all relatively prime positive integers n_1, n_2, $f(n_1 n_2) = f(n_1) \cdot f(n_2)$.

LEMMA 8.8. *If f is multiplicative, then $f(1) = 1$.*

PROOF. Taking $n_1 = n_2 = 1$, we have, using (M2)

$$f(1) = f(1 \cdot 1) = f(1) \cdot f(1) = f(1)^2.$$

Now by (M1), $f(1) \neq 0$, so that we may cancel $f(1)$'s to get $f(1) = 1$. □

Exercise: Suppose an arithmetic function f satisfies (M2) but not (M1). Show that $f \equiv 0$: i.e., $f(n) = 0$ for all $n \in \mathbb{Z}^+$.

The following is a nice characterization of multiplicative functions:

PROPOSITION 8.9. *For an arithmetic function f, the following are equivalent:*
a) f is multiplicative;
b) f is not identically zero, and for all $n = p_1^{a_1} \cdots p_k^{a_k}$ (the standard form factorization of n), we have $f(n) = \prod_{i=1}^k f(p_i^{a_i})$.

Remark: Here we are using the convention that for $n = 1$, $k = 0$, and a product extending over zero terms is automatically equal to 1 (just as a sum extending over zero terms is automatically equal to 0). (If this is not to your taste, just insert in part b) the condition that $f(1) = 1$!)

PROOF. Exercise. □

In other words, a multiplicative function f is completely determined by the values it takes on all prime powers p^k. Thus, in trying to understand a function known to be multiplicative, one needs only to "see what it is" on prime power values of n. Note that, conversely, any function f defined only on prime powers p^k – and satisfying $f(1) = 1$ – extends to a unique multiplicative function.

2.2. Completely multiplicative functions. If you have never seen this definition before, then something has been bothering you the whole time, and I will now respond to this worry. Namely, wouldn't it make more sense to say that a function f is multiplicative if $f(n_1 \cdots n_2) = f(n_1) \cdot f(n_2)$ for *all* integers n_1 and n_2?

In a purely algebraic sense the answer is yes: the stronger condition (together with $f(1) \neq 0$) says precisely that f is a homomorphism from the monoid $\mathbb{Z}^+$ to the monoid $\mathbb{C}^\times$. This is certainly a very nice property for f to have, and it has a name as well: **complete multiplicativity**. But in practice complete multiplicativity is *too nice*: only very special functions satisfy this property, whereas the class of multiplicative functions is large enough to contain many of our "arithmeticly significant functions." For instance, neither σ_k (for any k) nor φ is completely multiplicative, but, as we are about to see, all of these functions are multiplicative.

2.3. Multiplicativity of the σ_k's.

THEOREM 8.10. *The functions σ_k (for all $k \in \mathbb{N}$) are multiplicative.*

PROOF. It is almost obvious that the Möbius function is multiplicative. Indeed its value at a prime power p^a is: 1 if $a = 0$, -1 if $a = 1$, and 0 if $a \geq 2$. Now there is a unique multiplicative function with these values, and it is easy to see that μ is that function: we have $\mu(p_1^{a_1} \cdots p_k^{a_k}) = 0$ unless $a_i = 1$ for all i – as we should – and otherwise $\mu(p_1 \cdots p_k) = (-1)^k = \mu(p_1) \cdots \mu(p_k)$. In other words, μ is essentially multiplicative by construction.

Now let us see that σ_k is multiplicative. Observe that – since $\gcd(n_1, n_2) = 1$! – every divisor d of $n_1 \cdot n_2$ can be expressed uniquely as a product $d_1 \cdot d_2$ with

$d_i \mid n_i$. So

$$\sigma_k(n_1n_2) = \sum_{d \mid n_1n_2} d^k = \sum_{d_1 \mid n_1,\ d_2 \mid n_2} (d_1d_2)^k = (\sum_{d_1 \mid n_1} d_1^k)(\sum_{d_2 \mid n_2} d_2^k) = \sigma_k(n_1)\sigma_k(n_2).$$

□

2.4. CRT and the multiplicativity of the totient. The multiplicativity of φ is closely connected to the Chinese Remainder Theorem, as we now review. Namely, for coprime n_1 and n_2, consider the map $\Phi : \mathbb{Z}/(n_1n_2) \to \mathbb{Z}/(n_1) \times \mathbb{Z}/(n_2)$ given by

$$k \pmod{n_1n_2} \mapsto (k \pmod{n_1}, k \pmod{n_2}).$$

This map is a well-defined homomorphism of rings, since if $k_1 \equiv k_2 \pmod{n)_i}$, then $k_1 \equiv k_2 \pmod{n)_1n_2}$. Because the source and target have the same, finite, cardinality n_1n_2, in order for it to be an isomorphism it suffices to show either that it is injective or that it is surjective. Note that the standard, elementary form of the Chinese Remainder Theorem addresses the surjectivity: given any pair of congruence classes $i \pmod{n_1}$and $j \pmod{n_2}$ the standard proof provides an explicit formula for a class $p(i,j) \pmod{n_1n_2}$ which maps via Φ onto this pair of classes. However, writing down this formula requires at least a certain amount of cleverness, whereas it is trivial to show the injectivity: as usual, we need only show that the kernel is 0. Well, if $\Phi(\overline{k}) = 0$, then k is 0 mod n_1 and 0 mod n_2, meaning that $n_1 \mid k$ and $n_2 \mid k$. In other words, k is a common multiple of n_1 and n_2, so, as we've shown, k is a multiple of the least common multiple of n_1 and n_2. Since n_1 and n_2 are coprime, this means that $n_1n_2 \mid k$, i.e., that $k \equiv 0 \pmod{n_1n_2}$!

THEOREM 8.11. *There is a canonical isomorphism of groups*

$$(\mathbb{Z}/(n_1n_2))^\times \to (\mathbb{Z}/(n_1))^\times \times (\mathbb{Z}/(n_2))^\times.$$

PROOF. This follows from the isomorphism of rings discussed above, together with two almost immediate facts of pure algebra. First, if $\Phi : R \to S$ is an isomorphism of rings, then the restriction of Φ to the unit group $R^\times$ of R is an isomorphism onto the unit group $S^\times$ of S. Second, if $S = S_1 \times S_2$ is a product of rings, then $S^\times = S_1^\times \times S_2^\times$, i.e., the units of the product is the product of the units. We leave it to the reader to verify these two facts. □

COROLLARY 8.12. *The function φ is multiplicative.*

PROOF. Since $\varphi(n) = \#(\mathbb{Z}/(n))^\times$, this follows immediately. □

Now let us use the "philosophy of multiplicativity" to give exact formulas for $\sigma_k(n)$ and $\varphi(n)$. In other words, we have reduced to the case of evaluating at prime power values of n, but this is much easier. Indeed, the positive divisors of p^a are $1,\ p,\ \ldots, p^a$, so the sum of the kth powers of these divisors is $a+1$ when $k = 0$ and is otherwise

$$\sigma_k(p^a) = 1 + p^k + p^{2k} + \ldots + p^{ak} = \frac{1 - (p^k)^{a+1}}{1 - p^k} = \frac{1 - p^{(a+1)k}}{1 - p^k}.$$

Similarly, the only numbers $1 \leq i \leq p^a$ which are not coprime to p^a are the multiples of p, of which there are p^{a-1}: $1 \cdot p, 2 \cdot p, \ldots p^{a-1}p = p^a$. So

$$\varphi(p^a) = p^a - p^{a-1} = p^{a-1}(p-1) = p^a(1 - \frac{1}{p}).$$

COROLLARY 8.13. *Suppose* $n = p_1^{a_1} \cdots p_k^{a_k}$. *Then:*
a) $d(n) = \prod_{i=1}^k (a_i + 1)$.
b) For $k > 0$, $\sigma_k(n) = \prod_{i=1}^k \frac{1-p^{(a+1)k}}{1-p^k}$.
c) $\varphi(n) = \prod_{i=1}^k p^{a-1}(p-1)$.

The last formula is often rewritten as

$$\frac{\varphi(n)}{n} = \prod_{p \mid n} (1 - \frac{1}{p}). \tag{33}$$

While we are here, we quote the following more general form of the CRT, which is often useful:

THEOREM 8.14. *(Generalized Chinese Remainder Theorem) Let* $n_1, \ldots, n_r$ *be any* r *positive integers. Consider the natural map*

$$\Phi : \mathbb{Z} \to \mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z} \times \ldots \times \mathbb{Z}/n_r\mathbb{Z}$$

which sends an integer k *to* $(k \pmod{n_1}, \ldots, k \pmod{n_r})$.
a) The kernel of Φ *is the ideal* $(\mathrm{lcm}(n_1, \ldots, n_r))$.
b) The following are equivalent:
(i) Φ *is surjective;*
(ii) $\mathrm{lcm}(n_1, \ldots, n_r) = n_1 \cdots n_r$.
(iii) The integers $n_1, \ldots, n_r$ *are pairwise relatively prime.*

The proof is a good exercise. In fact the result holds essentially verbatim for elements $x_1, \ldots, x_r$ in a PID R, and, in some form, in more general commutative rings.

2.5. Additive functions. The function ω is not multiplicative: e.g. $\omega(1) = 0$ and $\omega(2) = 1$. However it satisfies a property which is "just as good" as multiplicativity: $\omega(n_1 n_2) = \omega(n_1) + \omega(n_2)$ when $\gcd(n_1, n_2) = 1$. Such functions are called **additive**. Finally, we have the notion of **complete additivity**: $f(n_1 n_2) = f(n_1) + f(n_2)$ for all $n_1,\ n_2 \in \mathbb{Z}^+$; i.e., f is a homomorphism from the positive integers under multiplication to the complex numbers under addition. We have seen some completely additive functions, namely, ord_p for a prime p.

PROPOSITION 8.15. *Fix any real number* $a > 1$ *(e.g.* $a = e,\ a = 2$*). A function* f *is additive (respectively, completely additive) iff* a^f *is multiplicative (respectively, completely multiplicative).*

PROOF. Exercise. □

2.6. Sums of squares. The functions r_k are not multiplicative: to represent 1 as a sum of k squares we must take all but one of the x_i equal to 0 and the other equal to ± 1. This amounts to $r_k(1) = 2k > 1$. However, we should not give up so easily! Put $r_k' = \frac{r_k}{2k}$. We can now quote a beautiful theorem, parts of which may be proved later.

THEOREM 8.16. *The function* r_k' *is multiplicative iff* $k = 1,\ 2,\ 4$ *or* 8.

2.7. Perfect numbers. The ancient Greeks regarded a positive integer n as **perfect** if it is equal to the sum of its proper divisors ("aliquot parts"). They knew some examples, e.g. 6, 28, 496, 8128.

In modern language a perfect number is a solution n of the equation $\sigma(n) - n = n$, or $\sigma(n) = 2n$. Aha, but we have an exact formula for the σ function: perhaps we can use it to write down all perfect numbers? The answer is a resounding "sort of."

Best, as usual, is to examine some data to try to figure out what is going on. Since our formula for σ takes into account the standard form factorization of n, we should probably look at these factorizations of our sample perfect numbers. We find:

$$6 = 2 \cdot 3$$
$$28 = 2^2 \cdot 7$$
$$496 = 2^4 \cdot 31$$
$$8128 = 2^6 \cdot 127.$$

As exercises in pattern recognition go, this is a pretty easy one. We have a power of 2 multiplied by an odd prime. But not just any odd prime, mind you, an odd prime which happens to be exactly one less than a power of 2. And not just any power of 2...anyway, we soon guess the pattern $2^{n-1} \cdot 2^n - 1$. But we're still not done: in our first four examples, n was $2, 3, 5, 7$, all primes. Finally we have a precise conjecture that our knowledge of σ can help us prove:

PROPOSITION 8.17. *(Euclid) Let p be a prime such that $2^p - 1$ is also prime. Then $N_p = 2^{p-1}(2^p - 1)$ is a perfect number.*

PROOF. Since $2^p - 1$ is odd, it is coprime to 2^{p-1}. So

$$\sigma(N_p) = \sigma(2^{p-1}(2^p - 1)) = \sigma(2^{p-1})\sigma(2^p - 1).$$

But these are both prime power arguments, so are easy to evaluate, as above. We get $\sigma(2^{p-1}) = 2^p - 1$ and $\sigma(2^p - 1) = 2^p$, so overall $\sigma(N_p) = 2^p \cdot (2^p - 1) = 2N_p$. □

This is a nice little calculation, but it raises more questions than it answers. The first question is: are there infinitely many primes p such that $2^p - 1$ is prime? Such primes are called **Mersenne primes** after Father Marin Mersenne, a penpal of Fermat. It would be appropriate to make any number of historical remarks about Mersenne and/or his primes, but we refer the reader to Wikipedia for this. Suffice it to say that, in theory, it is a wide open problem to show that there exist infinitely many Mersenne primes, but in practice, we do keep finding successively larger Mersenne primes (at a rate of several a year), meaning that new and ridiculously large perfect numbers are being discovered all the time.

Ah, but the second question: is every perfect number of the form N_p for a Mersenne prime p? Euler was able to show that every **even** perfect number is of this form. The argument is a well-known one (and is found in Silverman's book) so we omit it here. Whether or not there exist any odd perfect numbers is one of the notorious open problems of number theory. At least you should not go searching for odd perfect numbers by hand: it is known that there are no odd perfect numbers $N < 10^{300}$, and that any odd perfect number must satisfy a slew of restrictive conditions (e.g. on the shape of its standard form factorization).

3. Divisor Sums, Convolution and Möbius Inversion

The proof of the multiplicativity of the functions σ_k, easy though it was, actually establishes a more general result. Namely, suppose that f is a multiplicative function, and define a new function $F = \sum_d f$ as

$$F(n) = \sum_{d \mid n} F(d).$$

For instance, if we start with the function $f(n) = n^k$, then $F = \sigma_k$. Note that $f(n) = n^k$ is (in fact completely) multiplicative. The generalization of the proof is then the following

PROPOSITION 8.18. *If f is a multiplicative function, then so is $F(n) = \sum_{d \mid n} f(d)$.*

PROOF. If n_1 and n_2 are coprime, then $F(n_1 n_2) = \sum_{d | n_1 n_2} F(d) =$

$$\sum_{d_1 | n_1,\ d_2 | n_2} f(d_1 d_2) = \sum_{d_1 | n_1,\ d_2 | n_2} f(d_1) f(d_2) = (\sum_{d_1 | n_1} f(d_1))(\sum_{d_2 | n_2} f(d_2)) = F(n_1) F(n_2).$$

□

Exercise: Show by example that f completely multiplicative need not imply F completely multiplicative.

It turns out that the operation $f \mapsto F$ is of general interest; it gives rise to a certain kind of "duality" among arithmetic functions. Slightly less vaguely, sometimes f is simple and F is more complicated, but sometimes the reverse takes place.

Definition: Define the function δ by $\delta(1) = 1$ and $\delta(n) = 0$ for all $n > 1$. Note that δ is multiplicative. Also write ι for the function $n \mapsto n$.

PROPOSITION 8.19. *a) For all $n > 1$, $\sum_{d \mid n} \mu(d) = 0$.*
b) For all $n \in \mathbb{Z}^+$, $\sum_{d \mid n} \varphi(n) = n$.

In other words, the sum over the divisors of the Möbius function is δ, and the sum over the divisors of φ is ι.

PROOF. a) Write $n = p_1^{a_1} \cdots p_r^{a_r}$. Then $\sum_{d \mid} \mu(n) = \sum_{(\epsilon_1, \ldots, \epsilon_r)} \mu(p_1^{\epsilon_1} \cdots p_r^{\epsilon_r})$, where the ϵ_i are 0 or 1. Thus

$$\sum_{d \mid n} \mu(d) = 1 - r + \binom{r}{2} - \binom{r}{3} + \ldots + (-1)^r \binom{r}{r} = (1-1)^r = 0.$$

For part b) we take advantage of the fact that since φ is multiplicative, so is the sum over its divisors. Therefore it is enough to verify the identity for a prime power p^a, and as usual this is significantly easier:

$$\sum_{d \mid p^a} \varphi(p^a) = \sum_{i=0}^{a} \varphi(p^i) = 1 + \sum_{i=1}^{a} (p^i - p^{i-1}) = 1 + (p^a - 1) = p^a,$$

where we have cancelled out a telescoping sum. □

This indicates that the Möbius function is of some interest. We can go further by asking the question: suppose that $F = \sum_d f$ is multiplicative; must f be multiplicative?

Well, the first question is to what extent f is determined by its divisor sum function: if $F = \sum_d f = \sum_d g = G$, must $f = g$? If so, is there a nice formula which gives f in terms of F?

Some calculations:

$f(1) = F(1)$;

for any prime p, $F(p) = f(1) + f(p)$, so $f(p) = F(p) - F(1)$;

$F(p^2) = f(1) + f(p) + f(p^2) = F(p) + f(p^2)$, so $f(p) = F(p^2) - F(p)$; indeed

$f(p^n) = F(p^n) - F(p^{n-1})$.

For distinct primes p_1, p_2, we have $F(p_1p_2) = f(1) + f(p_1) + f(p_2) + f(p_1p_2)$
$= F(1) + F(p_1) - F(1) + F(p_2) - F(1) + f(p_1p_2)$, so

$$f(p_1p_2) = F(p_1p_2) - F(p_1) - F(p_2) + F(1).$$

This is an enlightening calculation on several accounts; on the one hand, there is some sort of inclusion-exclusion principle at work. On the other hand, and easier to enunciate, we are recovering f in terms of F and μ:

THEOREM 8.20. *(Möbius Inversion Formula) For any arithmetic function f, let $F(n) = \sum_{d|n} f(d)$. Then for all n,*

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

It is a good exercise to give a direct proof of this. However, playing on a familiar theme, we will introduce a little more algebra to get an easier proof. Namely, we can usefully generalize the construction $f \mapsto \sum_d f = F$ as follows:

Definition: For arithmetic functions f and g, we define their **convolution**, or **Dirichlet product**, as

$$(f * g)(n) = \sum_{d \mid n} f(d)g(\frac{n}{d}).$$

Why is this relevant? Well, define $\mathbf{1}$ as the function $\mathbf{1}(n) = 1$ for all n;[1] then $F = f * \mathbf{1}$. We have also seen that

$$\mu * \mathbf{1} = \iota, \tag{34}$$

and the inversion formula we want is

$$(f * \mathbf{1}) * \mu = f.$$

[1]We now have three similar-looking but different functions floating around: δ, ι and $\mathbf{1}$. It may help the reader to keep on hand a short "cheat sheet" with the definitions of all three functions.

Thus we see that if only it is permissible to rewrite $(f * \mathbf{1}) * \mu = f * (\mathbf{1} * \mu)$, then the inversion formula is an immediate consequence of Equation (34). In other words, we need to show that convolution is associative. In fact we can prove more:

PROPOSITION 8.21. *The arithmetic functions form a commutative ring under pointwise addition – i.e., $(f + g)(n) = f(n) + g(n)$ – and convolution. The multiplicative identity is the function δ.*

PROOF. In other words, we are making the following assertions: for all arithmetic functions f, g, h:
(i) $f * g = g * f$.
(ii) $(f * g) * h = f * (g * h)$.
(iii) $f * \delta = f$.
(iv) $f * (g + h) = f * g + f * h$.

To show both (i) and (ii) it is convenient to rewrite the convolution in symmetric form:

$$f * g(n) = \sum_{d_1 d_2 = n} f(d_1) g(d_2).$$

The sum extends over all pairs of positive integers d_1, d_2 whose product is n. This already makes the commutativity clear. As for the associativity, writing things out one finds that both $(f * g) * h$ and $f * (g * h)$ are equal to

$$\sum_{d_1 d_2 d_3 = n} f(d_1) g(d_2) h(d_3),$$

and hence they are equal to each other! For (iii), we have

$$(f * \delta)(n) = \sum_{d_1 d_2 = n} f(d_1) \delta(d_2);$$

$\delta(d_2) = 0$ unless $d_2 = 1$, so the sum reduces to $f(n)\delta(1) = f(n)$. The distributivity is easy and left to the reader. □

We can now show that $F = \sum_d f$ multiplicative implies f is multiplicative. Indeed, this follows from $f = F * \mu$, the multiplicativity of μ and the following:

PROPOSITION 8.22. *If f and g are multiplicative, so is $f * g$.*

PROOF. Just do it: for coprime m and n, $(f * g)(m)(f * g)(n) =$

$$\Big(\sum_{a_1 a_2 = m} f(a_1) g(a_2)\Big)\Big(\sum_{b_1 b_2 = n} f(b_1) g(b_2)\Big) = \sum_{a_1 a_2 b_1 b_2 = mn} f(a_1) f(b_1) g(a_2) g(b_2) =$$

$$\sum_{xy = mn} f(x) g(y) = (f * g)(mn).$$

□

4. Some Applications of Möbius Inversion

4.1. Application: another proof of the multiplicativity of the totient. Our first application of Möbius inversion is to give a proof of the multiplicativity of φ which is independent of the Chinese Remainder Theorem. To do this, we will

give a direct proof of the identity $\sum_{d|n} \varphi(d) = n$. Note that it is equivalent to write the left hand side as

$$\sum_{d|n} \varphi(\frac{n}{d}),$$

since as d runs through all the divisors of n, so does $\frac{n}{d}$.[2] Now let us classify elements of $\{1, \ldots, n\}$ according to their greatest common divisor with n. The greatest common divisor of any such element k is a divisor d of n, and these are exactly the elements k such that $\frac{k}{d}$ is relatively prime to $\frac{n}{d}$, or, in yet other words, the elements $d \cdot l$ with $1 \leq l \leq \frac{n}{d}$ and $\gcd(l, \frac{n}{d}) = 1$, of which there are $\varphi(\frac{n}{d})$. This proves the identity! Now, we can apply Möbius inversion to conclude that $\varphi = \iota \cdot \mu$ is multiplicative.

Here is a closely related approach. Consider the additive group of $\mathbb{Z}/n\mathbb{Z}$, a cyclic group of order n. For a given positive integer d, how many order d elements does it have? Well, by Lagrange's Theorem we need $d|n$. An easier question is how many elements there are of order dividing a given d (itself a divisor of n): these are just the elements $x \in \mathbb{Z}/n\mathbb{Z}$ for which $dx = 0$, i.e., the multiples of n/d, of which there are clearly d. But Möbius Inversion lets us pass from the easier question to the harder question: indeed, define $f(k)$ to be the number of elements of order k in $\mathbb{Z}/n\mathbb{Z}$; then $F(k) = \sum_{d|k} f(d)$ is the number of elements of order dividing k, so we just saw that $F(k) = k$. Applying Möbius inversion, we get that $f(k) = (I * \mu)(k) = \varphi(k)$. On the other hand, it is not hard to see directly that $f(k) = 0$ if k does not divide n and otherwise equals $\varphi(k)$ – e.g., using the fact that there is a unique subgroup of order k for all $k \mid n$ – and this gives another proof that $\varphi * \mathbf{1} = \iota$.

4.2. A formula for the cyclotomic polynomials. For a positive integer d, let $\Phi_d(x)$ be the monic polynomial whose roots are the primitive dth roots of unity, i.e., those complex numbers which have exact order d in the multiplicative group $\mathbb{C}^\times$ (meaning that $z^d = 1$ and $z^n \neq 1$ for any integer $0 < n < d$). These primitive roots are contained in the group of all dth roots of unity, which is cyclic of order d, so by the above discussion there are exactly $\varphi(d)$ of them: in other words, the degree of the polynomial Φ_d is $\varphi(d)$.[3] It turns out that these important polynomials have entirely integer coefficients, although without a somewhat more sophisticated algebraic background this may well not be so obvious. One might think that to write down formulas for the Φ_d one would have to do a lot of arithmetic with complex numbers, but that is not at all the case. Very much in the spirit of the group-theoretical interpretation of $\sum_{d|n} \varphi(d) = n$, we have

$$\prod_{d|n} \Phi_d(x) = x^n - 1,$$

since, both the left and right-hand sides are monic polynomials whose roots consist of each nth root of unity exactly once.

In fact it follows from this formula, by induction, that the Φ_d's have integral coefficients. But Möbius inversion gives us an explicit formula. The trick here is to convert the divisor product into a divisor sum by taking logarithms:

$$\log \prod_{d|n} \Phi_d(x) = \sum_{d|n} \log \Phi_d(x) = \log(x^n - 1).$$

[2]Alternately, this just says that $f * \mathbf{1} = \mathbf{1} * f$.

[3]For once, the ancients who fixed the notation have planned ahead!

Now applying Möbius inversion, we get

$$\log \Phi_n(x) = \sum_{d|n} \log(x^d - 1)\mu(\frac{n}{d})$$

$$= \log\left(\prod_{d|n}(x^d-1)^{\mu(n/d)}\right),$$

so exponentiating back we get a formula which at first looks too good to be true:

$$\Phi_n(x) = \prod_{d|n}(x^d-1)^{\mu(n/d)}.$$

But it is absolutely correct, and, as advertised, reduces the computation of the cyclotomic polynomials to arithmetic (including division!) of polynomials with integer coefficients.

Remark: Our trick of taking logs was done without much worry about rigor. It is not literally true that $\log(x^n-1)$ is an arithmetic function, since it is not defined for $x = 1$. We can justify what we have done as follows: for fixed n, since the Φ_d's and $x^n - 1$ are a finite set of monic polynomials with integer coefficients, there exists a large positive integer N such that for all d dividing n and all $x \in \mathbb{Z}^+$, $\Phi_d(x+N) \geq 1$, so that $\log(\Phi_d(x + N))$ and $\log(x + N)^d - 1$ are well-defined arithmetic functions, to which we apply MIF. This gives us the desired identity with $x + N$ in place of x, but being an identity of polynomials, we can substitute $x - N$ for x to get back to the desired identity.

On the other hand, such ad hockery is not so aesthetically pleasing. The formula we obtained suggests that MIF may be valid for functions f defined on $\mathbb{Z}^+$ but with more general codomains than $\mathbb{C}$. If R is a commutative ring, then the statement and proof of MIF go through verbatim for "R-valued arithmetic functions" $f : \mathbb{Z}^+ \to R$. But this is not the right generalization for the present example: we want a MIF for functions with values in the multiplicative group $\mathbb{C}(x)$ of nonzero rational functions. In fact, for any commutative group A – whose group law we will write as addition, even though in our application it is called multiplication – if one considers A-valued arithmetic functions $f : \mathbb{Z}^+ \to A$, then there is in general no convolution product (since we can't multiply elements of A), but nevertheless $F(n) = \sum_{d|n} f(d)$ makes sense, as does $\sum_{d|n} F(d)\mu(n/d)$, where for $a \in A$ we interpret $0 \cdot a$ as being the additive identity element 0_A, $1 \cdot a$ as a and $-1 \cdot a$ as the additive inverse $-a$ of a. Then one can check that $\sum_{d|n} F(d)\mu(n/d) = f(n)$ for all f, just as before. We leave the proof as an exercise.

4.3. Finite subgroups of unit groups of fields are cyclic.

THEOREM 8.23. *Let F be a field and $G \subset F^\times$ a finite subgroup of the multiplicative group of units of F. Then G is cyclic.*

PROOF. Suppose G has order n. Then, by Lagrange's theorem, we at least know that every element of G has order *dividing* n, and what we would like to know is that it has an element of order *exactly* n. We recognize this as an MIF situation, but this time MIF serves more as inspiration than a tool.

Namely, for any divisor d of n, let us define G_d to be the set of elements of

G of order dividing d. G_d is easily seen to be a subgroup of G, and subgroups of cyclic groups are cyclic, so if what we are trying to prove is true then G_d is a cyclic group of order d. Certainly this is true for $d = 1$! So assume by induction that this is true for all *proper* divisors d of n.

Now let $f(d)$ be the number of elements of order d in G. We know $f(d) = 0$ unless d is a divisor of n. We also know that for any d there are at most d elements in all of $F^\times$ of order d: the polynomial $x^d - 1$ can have at most d roots. Now suppose d is a proper divisor of n: our induction hypothesis implies that there are exactly d roots, namely the elements of G_d; moreover, since we are assuming that G_d is cyclic, of these d elements, exactly $\varphi(d)$ of them have exact order d. So $f(d) = \varphi(d)$ for all proper divisors d of n. But this means that the number of elements of G whose order is a proper divisor of n is $\sum_{d|n} \varphi(d) - \varphi(n) = n - \varphi(n)$, which leaves us with $n - (n - \varphi(n)) = \varphi(n)$ of elements of a group of order n whose order is not any proper divisor of n. The only possibility is that these elements all have order n, which is what we wanted to show. □

4.4. Counting irreducible polynomials.

Here is a truly classic application of Möbius Inversion.

THEOREM 8.24. *For any prime number p and any $n \in \mathbb{Z}^+$, the number of polynomials $P(x) \in \mathbb{Z}/p\mathbb{Z}[x]$ which are irreducible of degree n is*

$$I(\mathbb{Z}/p\mathbb{Z}, n) = \frac{1}{n}\left(\sum_{d|n} p^d \mu\left(\frac{n}{d}\right)\right).$$

The proof of Theorem 8.24 requires some preliminaries on polynomials over finite fields. We give a complete treatment in Appendix C.

5. A Bigger Möbius Inversion Formula

Our point of departure for the entire text was the relationship between arithmetic structure and order structure. We now revisit these concepts at a deeper level, obtaining a generalization of Möbius Inversion in suitable partially ordered sets. As a byproduct, we clarify the relationship between Möbius Inversion and the combinatorial principle of inclusion-exclusion. Our treatment follows [**BG75**] and [**St**].

5.1. Some Inversion Problems.

We begin with several examples that illustrate the notion of an "inversion problem."

EXAMPLE 8.25. *We begin with the Möbius Inversion that we already have: for a function $f : \mathbb{Z}^+ \to \mathbb{C}$, if we put*

$$F : \mathbb{Z}^+ \to \mathbb{C},\ F(n) \coloneqq \sum_{d|n} f(d),$$

then

$$\forall n \in \mathbb{Z}^+,\ f(n) = \sum_{d|n} F(d)\mu(n/d).$$

EXAMPLE 8.26. *(Calculus of Finite Differences) For a function $f : \mathbb{N} \to \mathbb{C}$, we define the* ***summatory function***

$$F : \mathbb{N} \to \mathbb{C},\ F(n) \coloneqq f(0) + f(1) + \ldots + f(n).$$

Question: Can we recover f from F?
Answer: Yes, quite easily: we have

$$f(0) = F(0),$$

$$\forall n \geq 1, f(n) = F(n) - F(n-1).$$

(If we take the convention that $F(-1) = 0$, then $f(0) = F(0) - F(-1)$.)
The operator Δ that takes a function g to the function $(\Delta g) : n \mapsto g(n) - g(n-1)$ is called the ***difference operator****. The operator that takes a function g to the function $\Sigma g : n \mapsto \sum_{i=0}^{n} g(n)$ is called the* ***summation operator****. Above we verified that for any $g : \mathbb{N} \to \mathbb{C}$, the relation*

$$\Delta(\Sigma(g)) = g.$$

It is similarly easy to verify the relation

$$\Sigma(\Delta(g)) = g :$$

indeed, recalling our convention that $g(-1) = 0$, we get

$$\Sigma(\Delta(g)) = (g(0)-g(-1))+(g(1)-g(0))+\ldots+(g(n-1)-g(n-2))+(g(n)-g(n-1)) = g(n).$$

One can take the perspective that the operator Δ is the discrete analogue of the derivative and that the operator Σ is the discrete analogue of the antiderivative. (Since antiderivatives are only unique up to constant, it is morally but not literally true that differentiation and anti-differentiation are mutually inverse operators. Here it is literally true.) This may not seem so auspicious, but in fact it is the jumping off point for a useful subfield of mathematics, the ***calculus of finite differences****.*

EXAMPLE 8.27. *(Inclusion-Exclusion, v1) We begin with the most classical statement of Inclusion-Exclusion. Let X be a finite set, and let $\{P_i\}_{i=1}^n$ be a family of subsets of X such that $\bigcup_{i=1}^n P_i = X$. Then there is a formula for the size of X in terms of the sizes of the P_i's and their intersections – namely,*

$$\#X = \sum_{i=1}^{n} \#P_i - \sum_{i \neq j} \#(P_i \cap P_j) + \sum_{distinct\ i,j,k} \#(P_i \cap P_j \cap P_j) - \ldots, \tag{35}$$

namely, the kth term is $(-1)^k$ times the sum of the cardinalities of all the k-fold intersections of the sets $P_1, \ldots, P_n$. This well known result can be proved by careful book-keeping; we urge the reader who has never seen this extremely useful formula before to give it a try. Let us first reformulate this as an inversion problem and then give a proof.

EXAMPLE 8.28. *(Inclusion-Exclusion, v2) As above, let X be a finite set, and let $\{P_i\}_{i=1}^n$ be a family of subsets of X. For each subset $T \subset \{1, \ldots, n\}$, we put*

$$N_+(T) \coloneqq \# \bigcap_{i \in T} P_i$$

and

$$N_=(T) \coloneqq \# \bigcap_{i \in T} P_i \cap \bigcap_{i \in \{1,\ldots,n\} \setminus T} (X \setminus P_i).$$

Let us give an interpretation: we may view each P_i as a "property" that an element of X may or may not have. Then $N_+(T)$ counts the nunmber of elements of X that have at least *the properties in the subset T (and possibly others), while $N_=(T)$ counts the number of elements of X that have* exactly *the properties in T. In many practical applications, $N_+(T)$ is easier to compute while $N_=(T)$ is of more interest. Clearly we have*

$$N_+(T) = \sum_{S \supset T} N_=(S),$$

so now we have an inversion problem: how to recover $N_=$ from N_+? The following formula does the trick:

$$N_=(T) = \sum -S \supset T (-1)^{\#(S \setminus T)} N_+(S). \tag{36}$$

In particular, taking $T = \varnothing$, we get a formula for the nunmber of elements of X that have "none of the properties":

$$N_= \varnothing = \sum_{S \subset \{1,\dots,n\}} (-1)^{\#S} N_+(S). \tag{37}$$

In (37), the first term on the right hand side is $(-1)^{\#\varnothing} N_+(\varnothing)$. This is the number of elements of X that have "at least none" of the properties, so it is $\#X$. Rearranging, we get

$$\#X - N_=(\varnothing) = \sum_{S \neq \varnothing} (-1)^{\#S+1} N_+(S).$$

But $\# - N_=(\varnothing)$ is precisely the set of elements of X that have at least one of the properties, so we get

$$\# \bigcup_{i=1}^{n} P_i = \sum_{\varnothing \neq S \subset \{1,\dots,n\}} (-1)^{\#S+1} \# \bigcap_{i \in S} P_i.$$

If we now add back the assumption that $\bigcup_{i=1}^n P_i = X$ as in the previous example, we recover (35).

Let us now give a proof of (36) – again, just using careful bookkeeping. Let $K = \#(\{1,\dots,n\} \setminus T)$, and for $0 \leq i \leq K$, let N_i be the set of elements $x \in N_+(T)$ such that $\#\{k \in \{1,\dots,n\} \setminus T \mid x \in P_k\} = i$ – in other words, that in addition to all the properties in T, x possesses precisely i further properties. Thus

$$N_+(T) = \coprod_{i=0}^{K} \#N_i \text{ and } \#N_0 = N_=(T).$$

Suppose zeroth that $x \in N_0$. Then x contributes 1 to $N_+(T)$ and nothing to $N_+(T)$ for all $S \supsetneq T$, so the total contribution to the right hand side of (37) from elements of N_0 is $\#N_0 = N_=(T)$.
Suppose first that $x \in N_1$. Thus there is a unique $i \in \{1,\dots,n\} \setminus S$ such that $x \in P_i$. Such an x contributes 1 to $N_+(S)$ and 1 to $N_+(S \cup \{i\})$, so its total contribution to the right hand side is $1 - 1 = 0$.
Suppose second that $x \in N_2$. Thus there are exactly two elements $i, j \in \{1,\dots,n\} \setminus S$ such that $x \in P_i$ and $x \in P_j$. Such an x contributes 1 to $N_+(S)$, 1 to each of $N_+(S \cup \{i\})$ and $N_+(S \cup \{j\})$ and 1 to $N_+(S \cup \{i,j\})$. The total contribution to the right hand side is $1 - 2 + 1 = 0$.

Now let $1 \leq i \leq K$. Then if $x \in S_i$ it lies in precisely i elements of $\{1, \ldots, n\} \setminus T$, so its total contribution to the right hand side of (37) is

$$\binom{K}{1} - \binom{K}{2} + \ldots + (-1)^K \binom{K}{K} = (1-1)^K = 0.$$

At this point we have precisely accounted for all the terms on the right hand side, so it evaluates to $\#N_0 + 0 + 0 + \ldots + 0 - \#N_0 = N_=(S)$.

EXAMPLE 8.29. *Let q be a prime power, let $n \in \mathbb{Z}^+$ and consider $\mathbb{F}_q^n$ as an n-dimensional vector space over the finite field $\mathbb{F}_q$ of order q. For any $\mathbb{F}_q$-linear subspace $U \subset \mathbb{F}_q^n$, let $N_=(U)$ be the number of subsets $S \subset \mathbb{F}_q^n$ whose span is U. Also let $N_\leq(U)$ be the number of subsets $S \subset \mathbb{F}_q^n$ whose span is contained in U. Then we have*

$$N_\leq(U) = \sum_{V \subset U} N_=(V), \tag{38}$$

where the sum ranges over all subspaces V of U. But actually $N_\leq(U)$ is trivial to compute: a subset S of a vector space has span lying in that subspace iff the set lies in the subspace, so $N_\leq(U)$ is just $2^{\#U}$, the number of subsets of U. Writing $\dim U$ for the dimension of U, we have $\#U = q^{\dim U}$ and thus

$$N_\leq(U) = 2^{q^{\dim U}}.$$

The quantity $N_=(U)$ is a bit more interesting, so we would like to somehow "invert" (38) to get a formula for the $N_=$'s in terms of the $N_\leq$'s. This time we leave the problem unsolved for now and turn to the development of the general theory.

5.2. The Möbius Function on a Locally Finite Poset.

The above examples fit into a common framework. Recall that a partially ordered set, or "poset," is a set X equipped with a binary operation $\leq$ that is reflexive, symmetric and transitive. A poset is **totally ordered** (or **linearly ordered**, or a **chain**) if for all $x, y \in X$, either $x \leq y$ or $y \leq x$. A **bottom element** of a poset is an element B such that $B \leq x$ for all x in X. A **top element** of a poset is an element T such that $x \leq T$ for all x in X. (A poset has either one bottom element or none at all; the same goes for top elements.) For every poset $(X, \leq)$ we define the **dual** poset $X^\vee$: it has the same underlying set and the transposed order relation: that is,

$$x \leq_\vee y \iff y \leq x.$$

Certain properties of a poset pair off via passage to the dual poset; we call such properties "dual." For instance, a top element is dual to a bottom element.

For $x \leq y$ in a poset X, we define the **inverval**

$$[x, y] \coloneqq \{z \in X \mid x \leq z \leq y\}.$$

We also define

$$(x, y] \coloneqq [x, y] \setminus \{x\},\ [x, y) \coloneqq [x, y] \setminus \{y\}.$$

For $x \leq y$ in X, we say that **y covers x** if $[x, y] = \{x, y\}$; in other words, if $x < y$ and there is no z with $x < z < y$.

A poset is **locally finite** if all its intervals are finite.

EXERCISE 8.1. Let $(X, \leq)$ be a poset. A **linear extension** of X is a total order relation $\leq'$ on X such that $\leq \subset \leq'$: that is, for all $x, y \in X$, if $x \leq y$ then also $x \leq' y$.
a) Show: every finite poset admits at least one linear extension.
(Suggestion: if X is a finite poset that is not totally ordered, there are elements x and y such that $x \not\leq y$ and $y \not\leq x$. Enlarge the relation to $\leq_1$ by putting $x \leq_1 y$. Then $\leq_1$ need not be a partial ordering, but its transitive closure is. Repeat until you get a total ordering.)
b) Show: every poset admits at least one linear extnesion.
(Suggestion: use Zorn's Lemma.)

EXERCISE 8.2. Let $(X, \leq)$ be a poset, and let $x \leq y$ be elements of X. A **covering chain from x to y** is a finite sequence $\{z_0, \ldots, z_n\}$ of elements of X such that z_{i+1} covers z_i for all $0 \leq i \leq n-1$, $z_0 = x$ and $z_n = y$. Show: if X is locally finite, then whenever $x \leq y$ there is a covering chain from x to y. Is the converse true?

EXERCISE 8.3.
a) Let $(X, \leq)$ be a poset and let Y be a subset of X. Then Y becomes a poset just by restricting the relation $\leq$. Show: if X is locally finite, then so is Y.
b) Show: a finite poset is locally finite.
c) Show: the integers $\mathbb{Z}$ under the usual ordering is locally finite.
d) Show: the positive integers $\mathbb{Z}^+$ under the divisibility ordering is locally finite.
e) If X is a set and $\{Y_i\}_{i \in I}$ is a family of subsets of X, inclusion gives a partial ordering on the index set I: we put $Y_i \leq Y_j$ iff $Y_i \subset Y_j$. An important special case is the family 2^X of all subsets of X. Show: 2^X is locally finite iff X is finite.
f) Let V be a vector space over a field F, and let $\mathrm{Sub}(V)$ be the family of all F-linear subspaces of V, partially ordered by inclusion. Show: $\mathrm{Sub}(V)$ is locally finite iff either $\dim V \leq 1$ or F and $\dim V$ are both finite.

EXERCISE 8.4. A poset is **downward finite** if for all $y \in X$, the set $D(y) := \{x \in X \mid x \leq y\}$ is finite. Dually, a poset is **upward finite** if for all $y \in X$, the set $U(y) := \{x \leq X \mid x \geq y\}$ is finite. Show: if X has a bottom element and is locally finite, then it is downward finite, and (dually) that if X has a top element and is locally finite, then it is upward finite.

Let $(X, \leq)$ be a locally finite poset. We claim there is a unique function $\mu : X \times X \to \mathbb{Z}$ satisfying the following properties:
• $\mu(x, y) = 0$ unless $x \leq y$.
• $\mu(x, x) = 1$ for all $x \in X$.
• if $x < y$, then $\mu(x, y) = -\sum_{z \in [x,y)} \mu(x, z)$.

To see this we argue by induction on the maximal length of a covering chain from x to y. If this length is 1 – i.e., if y covers x – we get $\mu(x, y) = -\mu(x, x) = -1$. Assuming that we have defined $\mu(x, y)$ when all covering chains from x to y have length at most L, suppose that the maximal length of a covering chain from x to y is $L + 1$. Then for all $z \in [x, y)$, every covering chain from x to z has length at most L, so $\mu(x, z)$ is already defined. So the formula $\mu(x, y) = -\sum_{z \in [x,y)} \mu(x, z)$ defines $\mu(x, y)$.

An easy but important remark: for elements $x \leq y$ in a locally finite poset, the value $\mu(x,y)$ depends only on the interval $[x,y]$, which is a finite poset.

EXAMPLE 8.30. *We compute the Möbius function on $\mathbb{Z}$. In fact for all $n \in \mathbb{Z}$ we have $\mu(n,n) = 1$ and $\mu(n,n+1) = -1$, and we have $\mu(m,n) = 0$ unless $n-m \in \{0,1\}$. If $m < n$, then $[m,n]$ is isomorphic as a poset to $[0,n-m]$, so we may as well look at the finite chain $0 < 1 < \ldots < N$, and in this case the result is virtually immediate, as we leave to the reader to verify.*

EXERCISE 8.5. Let X be a nonempty locally finite totally ordered set. Show: if X is finite then it is order isomorphic to the interval $[0, \#X-1]$ in the integers. If X is infinite, then it is isomorphic to either the non-negative integers, the non-positive integers, or the integers.

EXAMPLE 8.31. *We compute the Möbius function on $\mathbb{Z}^+$ with the divisibility ordering. Let $m,n \in \mathbb{Z}^+$ with $m \mid n$. Then the interval $[m,n]$ is order isomorphic to the interval $[1,\frac{n}{m}]$, so $\mu(m,n) = \mu(1,\frac{n}{m})$. We claim that in fact $\mu(1,N) = \mu(N)$ – where on the left hand side we have our new 2-variable Möbius function and on the right hand side we have the arithmetic function; this has the consequence that*

$$\mu(m,n) = \mu(\frac{n}{m}).$$

This is easy to see by induction on N: we have $\mu(1,1) = 1 = \mu(1)$; now let $N \geq 2$ and suppose that $\mu(1,m) = \mu(m)$ for all $1 \leq m < N$. Then we have

$$\mu(1,N) = -\sum_{y \in [1,N)} \mu(1,y) \stackrel{\text{IH}}{=} -\sum_{y|N,\ y<N} \mu(y) = \left(-\sum_{y|N} \mu(y)\right) + \mu(N) = \mu(N).$$

We want to comput the Möbius function on $2^{[1,n]}$, the set of all subsets of an n-element set. For this we want to make use of a "product structure." For posets $(X_1, \leq_1)$ and $(X_2, \leq_2$ we define a new poset as follows: the underlying set X is the Cartesian product $X_1 \times X_2$, and for $x = (x_1,x_2),\ y = (y_1,y_2) \in X$, we put

$$(x_1,x_2) \leq (y_1,y_2) \iff x_1 \leq y_1 \text{ and } x_2 \leq y_2.$$

This is natural and useful: e.g. it is a good way of producing non-total orders.

EXERCISE 8.6. Let X_1, X_2 be nonempty posets, and let $X = X_1 \times X_2$ with the product ordering. Show that X is totally ordered iff X_1 and X_2 are both totally ordered and ($\#X_1 = 1$ or $\#X_2 = 1$).

EXERCISE 8.7. Let X_1, X_2 be posets. For $x_1 \leq y_1 \in X_1$ and $x_2 \leq y_2 \in X_2$, put $x = (x_1,x_2),\ y = (y_1,y_2) \in X_1 \times X_2$.
a) Show: We have $[x,y] = [x_1,y_1] \times [x_2,y_2]$.
b) Deduce: if X_1 and X_2 are locally finite, so is $X_1 \times X_2$.

We can immediately extend this notion to an n-fold product (and in fact to the product of any indexed family of posets). Now we observe that the poset $2^{[1,n]}$ is order isomorphic to the poset $\prod_{i=1}^n [0,1]$ of length n binary strings. (Here one binary string is less than or equal to another if the ith digit of the first string is less than or equal to the ith digit of the second string for all $1 \leq i \leq n$.) Namely, to a subset $A \subset [1,n]$ we attach the "indicator string" i_A whose ith digit is 1 if $i \in A$ and 0 otherwise. This is an order-preserving bijection.

The point is that there is a simple expression for the Möbius function of a product.

THEOREM 8.32. *(Product Theorem) Let X_1, X_2 be locally finite posets, and put $X = X_1 \times X_2$ with the product ordering. Then we have*

$$\mu_X = \mu_{X_1} \times \mu_{X_2} : \forall (x_1, x_2),\ (y_1, y_2) \in X,\ \mu_X((x_1, x_2), (y_1, y_2)) = \mu_{X_1}(x_1, y_1)\mu_{X_2}(x_2, y_2).$$

PROOF. Let $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Put $M(x, y) \coloneqq \mu_{X_1}(x_1, y_1)\mu_{X_2}(x_2, y_2)$. We will show that M satisfies the three defining properties of the Möbius function.
• If $x \not\leq y$ then $x_1 \not\leq y_1$ – so $\mu_{X_1}(x_1, y_1) = 0$ – or $x_2 \not\leq y_2$ – so $\mu_{X_2}(x_2, y_2) = 0$. Either way we have $M(x, y) = \mu_{X_1}(x_1, y_1)\mu_{X_2}(x_2, y_2) = 0$.
• If $x = y$ then $M(x, y) = \mu_{X_1}(x_1, x_1)\mu_{X_2}(x_2, x_2) = 1 \cdot 1 = 1$.
• Let $x < y$. Then

$$\sum_{z \in [x,y]} M(x, z) = \sum_{(x_1,x_2) \leq (z_1,z_2) \leq (y_1,y_2)} \mu_{X_1}(x_1, z_1)\mu_{X_2}(x_2, z_2)$$

$$= \left(\sum_{z_1 \in [x_1,y_1]} \mu_{X_1}(x_1, z_1) \right) \left(\sum_{z_2 \in [x_2,y_2]} \mu_{X_2}(x_2, z_2) \right).$$

Since $x < y$, either $x_1 < y_1$ – in which case the first factor above is zero – or $x_2 < y_2$ – in which case the second factor above is zero. Either way the product is zero. □

EXAMPLE 8.33. *We return to $2^{[1,n]} \cong \{0, 1\}^n$. The Möbius function on $\{0, 1\}$ is certainly known to us: we have*

$$\mu(0, 0) = \mu(1, 1) = 1,\ \mu(0, 1) = -1,\ \mu(1, 0) = 0,$$

and thus for subsets $A \subset B \subset [1, n]$, we get

$$\mu(A, B) = (-1)^{\#B - \#A}.$$

EXAMPLE 8.34. *With the Product Theorem in hand, we revisit the ordinary Möbius function $\mu : \mathbb{Z}^+ \to \mathbb{Z}$. Let $n = p_1^{a_1} \cdots p_r^{a_r} \in \mathbb{Z}^+$. Let $D(n)$ be the set of positive divisors of n, partially ordered by divisibility. Then we have*

$$D(n) \cong \prod_{i=1}^{r} D(p_i^{a_i}),$$

and thus

$$\mu(n) = \mu(1, n) = \prod_{i=1}^{n} \mu(1, p_i^{a_i}).$$

Since we have already computed the Möbius function on a totally ordered set, we know that $\mu(1, p^{a_i}) = \begin{cases} -1 & a_i = 1 \\ 0 & a_i \geq 2 \end{cases}$. We conclude that $\mu(n)$ is 0 unless n is square-free in which case it is $(-1)^r$. Thus we get an interpretation of our definition of the classical Möbius function in terms of properties of the 2-variable Möbius function of a locally finite poset.

5.3. The Inversion Formula.

THEOREM 8.35. *(Möbius Inversion) Let X be a locally finite poset.*
a) Assume moreover that X is downward finite (this holds if X has a bottom element). For a function $f : X \to \mathbb{C}$, we define

$$F : X \to \mathbb{C},\ x \mapsto \sum_{y \leq x} f(y).$$

Then

$$\forall x \in X, f(x) = \sum_{y \leq x} f(y)\mu(y,x).$$

b) Assume moreover that X is upward finite (this holds if X has a top element). For a function $f : X \to \mathbb{C}$, we define

$$F : X \to \mathbb{C},\ x \mapsto \sum_{y \geq x} f(y).$$

Then

$$\forall x \in X, f(x) = \sum_{y \geq x} f(y)\mu(x,y).$$

The two parts of Theorem 8.35 are duals: applying part a) to $X^\vee$, we get part b).

We proved classical Möbius Inversion by introducing a convolution product on arithmetic functions and identiying the Möbius function as the inverse to the constant function $\mathbf{1}$. We will do something broadly similar here, but there must be some differences because in the context of Theorem 8.35 our Möbius function is not a function on X but rather on $X \times X$ – or, if we like, on the set $\mathrm{Int}(X)$ of intervals in X. So we are looking for a product operation on

$$I(X,\mathbb{C}) \coloneqq \{\text{functions } f : \mathrm{Int}(X) \to \mathbb{C}\},$$

which makes it into a $\mathbb{C}$-algebra. Observe that $I(X,\mathbb{C})$ is naturally a $\mathbb{C}$-vector space under pointwise addition and scalar multiplication. The following is the desired product: for $f, g : \mathrm{Int}(X) \to \mathbb{C}$, we put

$$f * g : [x,y] \mapsto \sum_{x \leq z \leq y} f([x,z])g([z,y]).$$

A nice way to think of this product is defined by thinking of $I(X,\mathbb{C})$ as the set of all infinite formal $\mathbb{C}$-linear combinations $\sum_{[x,y]} f(x,y)[x,y]$. We then put

$$[x,y] * [z,w] = \begin{cases} [x,w] & y = z \\ 0 & \text{otherwise} \end{cases}$$

and we extend by bilinearity (allowing infinite linear combinations). The unit element for this product is

$$\delta : \begin{cases} [x,x] \to 1 & \forall x \in X \\ [x,y] \to 0 & \forall x < y \in X. \end{cases}$$

Both $(f * g) * h$ and $f * (g * h)$ take $[x,y]$ to $\sum_{x \leq z \leq w \leq y} f([x,z])g([z,w])h([w,y])$, and thus the algebra is associative.

In the case that X is finite of cardinality n, there is a particularly concrete representation of $I(X,\mathbb{C})$, as follows: choose a linear extension of X; equivalently, write $X = \{x_1, \ldots, x_n\}$ so that $x_i \leq x_j$ if $i \leq j$. Then $I(X,\mathbb{C})$ can be identified with the subring of $n \times n$ complex matrices $M = (m_{ij})$ such that $m_{ij} = 0$ unless $x_i \leq x_j$: namely we map $f([x_i, x_j])$ to m_{ij}. This provides another way of showing the associativity in this case. Moreover it shows that $I(X,\mathbb{C})$ can be non-commutative: e.g. when $n \geq 2$ and X is totally ordered, then $I(X,\mathbb{C})$ can be identified with the ring of all upper triangular $n \times n$ matrices.

PROPOSITION 8.36. *For $f \in I(X, \mathbb{C})$, the following are equivalent:*
(i) f has a left inverse.
(ii) f has a right inverse.
(iiii) f has a two-sided inverse.
(iv) We have $f([x,x]) \neq 0$ for all $x \in X$.

PROOF. We have $fg = \delta$ iff for all $x, y \in X$ we have

(39)
$$f([x,x])g([x,x]) = 1,\ g([x,y]) = -f([x,x])^{-1} \sum_{z \in (x,y]} f([x,z])g([z,y]) \text{ if } x < y.$$

This shows that f has a right inverse iff $f([x,x]) \neq 0$ for all $x \in X$. Exactly the same reasoning shows that f has a right inverse iff $f([x,x]) \neq 0$ for all $x \in X$, so if $f([x,x]) \neq 0$ for all $x \in X$ then there is g_1 such that $f * g_1 = \delta$ and g_2 such that $g_2 * f = \delta$. But then

$$g_1 = \delta * g_1 = (g_2 * f) * g_1 = g_2 * (f * g_1) = g_2 * \delta = g_2,$$

so we have a two-sided inverse. □

Now we define the **zeta function** $\zeta : \mathrm{Int}(X) \to \mathbb{C}$ by

$$\zeta([x,y]) = 1\ \forall x \leq y \in X.$$

This is the analogue of the function **1** in the classical case.

EXERCISE 8.8. a) Show: $\zeta^2([x,y]) = \#[x,y]$.
b) Show: for $k \in \mathbb{Z}^+$, $\zeta^k([x,y]) = \sum_{x = x_0 \leq x_1 \leq \ldots \leq x_k = y} 1$.

By Proposition 8.36, the zeta function ζ has an inverse in $I(X, \mathbb{C})$. And, as expected, we have $\mu = \zeta^{-1}$: indeed if we apply (39) with $f = \zeta$, we see that g obeys the defining relations for the Möbius function. To write them in a more symmetric form, they are:

$$\mu([x,x]) = 1\ \forall x \in X,\ \sum_{x \leq z \leq y} \mu([x,z]) = 1\ \forall x < y \in X.$$

The Möbius algebra $I(X, \mathbb{C})$ has a natural representation on the $\mathbb{C}$-vector space $\mathbb{C}^X = \{f : X \to \mathbb{C}\}$ by: for $f \in \mathbb{C}^X$ and $E \in I(X, \mathbb{C})$, put

$$(f \circ E)(x) = \sum_{y \leq x} f(y)E([y,x]).$$

Here by "representation" we mean the fact that

$$((f \circ E_1) \circ E_2)(x) = \sum_{y \leq x} (f \circ E_1)(y)E_2([y,x]) = \sum_{y \leq x} \sum_{z \leq y} f(z)E_1[z,y]E_2[y,x]$$

$$= \sum_{z \leq x} f(z)(E_1 * E_2)[z,x] = f \circ (E_1 * E_2).$$

Thus, for $f, F \in \mathbb{C}^X$, we have

$$F = f \circ \zeta \iff f = F \circ \mu :$$

this is Theorem 8.35a).

EXERCISE 8.9. Let $(G, +)$ be a commutative group, and let $(X, \leq)$ be a downward finite poset. For a function $f : X \to G$, define

$$F : x \mapsto \sum_{y \leq x} f(y).$$

Show: for all $x \in X$, we have

$$f(x) = \sum_{y \leq x} F(y)\mu([y, x]).$$

(Suggestion: define the Möbius ring $I(X, \mathbb{Z})$ to be as for $I(X, \mathbb{C})$ but with $\mathbb{C}$ replaced by $\mathbb{Z}$. Show that we have μ and ζ in $I(X, \mathbb{Z})$ exactly as above. Moreover, show that $I(X, C)$ acts by $\mathbb{Z}$-linear endomorphisms of $\mathbb{Z}^G = \{f : G \to \mathbb{Z}\}$ and thereby deduce Möbius inversion exactly as above.

CHAPTER 9

Asymptotics of Arithmetic Functions

1. Introduction

Having entertained ourselves with some of the more elementary and then the more combinatorial/algebraic aspects of arithmetic functions, we now grapple with what is fundamentally an analytic number theory problem: for a given arithmetic function f, approximately how large is $f(n)$ as a function of n?

It may at first be surprising that this is a reasonable – and, in fact, vital – question to ask even for the "elementary" functions f for which we have found exact formulas, e.g. $d(n)$, $\sigma(n)$, $\varphi(n)$, $\mu(n)$ (and also $r_2(n)$, which we have not yet taken the time to write down a formula for but could have based upon our study of the Gaussian integers). What we are running up against is nothing less than the multiplicative/additive dichotomy that we introduced at the beginning of the course: for simple multiplicative functions f like d and φ, we found exact formulas. But these formulas were not directly in terms of n, but rather made reference to the standard form factorization $p_1^{a_1} \cdots p_r^{a_r}$. It is easy to see that the behavior of, say, $\varphi(n)$ as a function of "n alone" cannot be so simple. For instance, suppose $N = 2^p - 1$ is a Mersenne prime. Then

$$\varphi(N) = N - 1.$$

But

$$\varphi(N+1) = \varphi(2^p) = 2^p - 2^{p-1} = 2^{p-1} = \frac{N+1}{2}.$$

This is a bit disconcerting: $N+1$ is the tiniest bit larger than N, but $\varphi(N+1)$ is half the size of $\varphi(N)$!

Still we would like to say something about the size of $\varphi(N)$ for large N. For instance, we saw that for a prime p there are precisely $\varphi(p-1)$ primitive roots modulo p, and we would like to know something about how many this is.

Ideal in such a situation would be to have an asymptotic formula for φ: that is, a simple function $g : \mathbb{Z}^+ \to (0, \infty)$ such that $\lim_{n\to\infty} \frac{\varphi(n)}{g(n)} = 1$. (In such a situation we would write $\varphi \sim g$.) But it is easy to see that this is too much to ask. Indeed, as above we have $\varphi(p) = p - 1$, so that restricted to prime values $\varphi(p) \sim p$; on the other hand, restricted to even values of n, $\varphi(n) \leq \frac{n}{2}$, so there is too much variation in φ for there to be a simple asymptotic expression.

This is typical for the classical arithmetic functions; indeed, some of them, like the divisor function, have even worse behavior than φ. In other words, φ has more than one kind of limiting behavior, and there is more than one relevant question to ask. We may begin with the following:

QUESTION 4. *a) Does $\varphi(n)$ grow arbitrarily large as n does?*
b) How small can $\varphi(n)/n$ be for large n?

Part a) asks about the size of φ in an absolute sense, whereas part b) is asking about φ in a relative sense. In particular, since there are $\varphi(p) = p-1$ elements of $(\mathbb{Z}/p\mathbb{Z})^\times$, the quantity $\frac{\varphi(p-1)}{p-1}$ measures the chance that a randomly chosen nonzero residue class is a primitive root modulo p. Note we ask "how small" because we know how large $\frac{\varphi(n)}{n}$ can be: arbitrarily close to 1, when n is a large prime.

2. Lower bounds on Euler's totient function

Anyone who works long enough with the φ function (for instance, in computing all n such that $\varphi(n) \leq 10$) will guess the following result:

PROPOSITION 9.1. *We have* $\lim_{n\to\infty} \varphi(n) = \infty$.

Equivalently: for any $L \in \mathbb{Z}^+$, there are only finitely many n such that $\varphi(n) \leq L$.

The idea of the proof is a simple and sensible one: if a positive integer n is "large", it is either divisible by a large prime p, or it is divisible by a large power a of a prime, or both. To formalize this a bit, consider the set $S(A,B)$ of positive integers n which are divisible only by primes $p \leq A$ and such that $\mathrm{ord}_p(n) \leq B$ for all primes p. Then $S(A,B)$ is a finite set: indeed it has at most $(B+1)^A$ elements. (Also its largest element is at most $\prod_{p\leq A} p^B \leq (A!)^B$, which is, unfortunately, pretty darned large.)

So if we assume that n is sufficiently large – say larger than $(L!)^L$ – then n is divisible either by a prime $p > L$ or by p^{L+1} for some prime p. It is easy to show that if $m \mid n$, $\varphi(m)|\varphi(n)$ – and thus $\varphi(m) \leq \varphi(n)$. So in the first case we have

$$\varphi(n) \geq \varphi(p) = p - 1 \geq L,$$

and in the second case we have

$$\varphi(n) \geq \varphi(p^{L+1}) = p^L(p-1) \geq p^L > L.$$

So we've shown that if $n > (L!)^L$, then $\varphi(n) \geq L$, which proves the result.

It was nice to get an explicit lower bound on φ, but the bound we got is completely useless in practice: to compute all n for which $\varphi(n) \leq 5$ above argument tells us that it suffices to look at n up to $120^5 = 24883200000$. But this is ridiculous: *ad hoc* arguments do much better. For instance, if n is divisible by a prime $p \geq 7$, then $\varphi(n)$ is divisible by $p-1 \geq 6$, so we must have $n = 2^a3^b5^c$. If $c \geq 2$, then $25 \mid n$ so $20 = \varphi(25) \leq \varphi(n)$. Similarly, if $b \geq 2$, then $9 \mid n$ so $6 = \varphi(9) \leq \varphi(n)$, and if $a \geq 4$, then $16 \mid n$ so $8 = \varphi(16) \leq \varphi(n)$. So, if $n = 5m$, then $\varphi(n) = 4\varphi(m)$ so $\varphi(m) = 1$ and thus $m = 1$ or 2. If $n = 3m$, then $\varphi(n) = 2\varphi(m)$, so $\varphi(m) = 1$ or 2, so $n = 3 \cdot 1,\ 3 \cdot 2,\ 3 \cdot 4$. Otherwise n is not divisible by 9 or by any prime $p \leq 5$, so that $b \leq 1$ and $a \leq 3$. This yields the possibilities $n = 1, 2, 4, 8, 3, 6$. In summary, $\varphi(n) \leq 5$ iff

$$n = 1, 2, 3, 4, 5, 6, 8, 10, 12.$$

More practical lower bounds are coming up later.

However, it is interesting to note that essentially the same idea allows us to give us a much better asymptotic lower bound on φ. Namely, we have the following

pretty result which once again underscores the importance of keeping an eye out for multiplicativity:

THEOREM 9.2. *Suppose f is a multiplicative arithmetic function such that $f(p^a) \to 0$ as $p^a \to \infty$. Then $f(n) \to 0$ as $n \to \infty$.*

In other words if f is a multiplicative function such that for every $\epsilon > 0$, $|f(p^m)| < \epsilon$ for all sufficiently large prime powers, it follows that $|f(n)| < \epsilon$ for all sufficiently large n, prime power or otherwise.

Remark: As long as our multiplicative function f is never 0, an equivalent statement is that if $f(p^n) \to \infty$ for all prime powers than $f(n) \to \infty$ for all n. (Just apply the theorem to $g = \frac{1}{f}$, which is multiplicative iff f is.) So assuming the theorem, we can just look at

$$\varphi(p^a) = p^{a-1}(p-1) \geq \max(p-1, a-1),$$

and if p^a is large, at least one of p and a is large. But actually we get more:

COROLLARY 9.3. *For any fixed δ, $0 < \delta < 1$, we have $\varphi(n)/n^\delta \to \infty$.*

PROOF. We wish to show that $f(n) := \frac{n^\delta}{\varphi(n)} \to 0$ as $n \to \infty$. Since both n^δ and $\varphi(n)$ are multiplicative, so is their quotient f, so by the theorem it suffices to show that f approaches zero along prime powers. No problem:

$$f(p^n) = \frac{p^{n\delta}}{p^{n-1}(p-1)} = \frac{p}{p-1} \cdot (p^{\delta-1})^n.$$

Here $\delta - 1 < 0$, so as $p \to \infty$ the first factor approaches 1 and the second factor approaches 0 (just as $x^\alpha \to 0$ as $x \to \infty$ for negative α). On the other hand, if p stays bounded and $n \to \infty$ then the expression tends to 0 exponentially fast. □

Now let us prove Theorem 9.2. We again use the idea that for any $L > 0$, there exists $N = N(L)$ such that $n > N$ implies N is divisible by a prime power $p^a > L$.

First let's set things up: since $f(p^m) \to 0$ we have that f is bounded on prime powers, say $|f(p^m)| \leq C$. Moreover, there exists a b such that $|f(p^m)| \leq 1$ for all $p^m \geq b$; and finally, for every $\epsilon > 0$ there exists $L(\epsilon)$ such that $p^m > L(\epsilon)$ implies $|f(p^m)| < \epsilon$. Now write $n = p_1^{a_1} \cdots p_r^{a_r}$, so that

$$f(n) = f(p_1^{a_1}) \cdots f(p_r^{a_r}).$$

Since there are at most b indices i such that $p_i^{a_i} \leq B$, there are at most b factors in the product which are at least 1 in absolute value, so that the product over these "bad" indices has absolute value at most C^b. Every other factor has absolute value at most 1. Moreover, if n is sufficiently large with respect to $L(\epsilon)$ (explicitly, if $n > L(\epsilon)!^{L(\epsilon)}$, as above), then the largest prime power divisor $p_r^{a_r}$ of n is greater than $L(\epsilon)$ and hence $|f(p_r^{a_r})| < \epsilon$. This gives

$$|f(n)| = |f(p_1^{a_1} \cdots p_r^{a_r})| \leq C^b \cdot \epsilon.$$

Since C and b are fixed and ϵ is arbitrary, this shows that $f(n) \to 0$ as $n \to \infty$.

A nice feature of Theorem 9.2 is that it can be applied to other multiplicative functions. For instance, it allows for a quick proof of the following useful upper bound on the divisor function:

THEOREM 9.4. *For every fixed $\delta > 0$, we have*

$$\lim_{n\to\infty} \frac{d(n)}{n^\delta} = 0.$$

PROOF. Exercise. □

Note that Corollary 9.3 is equivalent to the following statement: for every $0 < \delta < 1$, there exists a positive constant $C(\delta)$ such that for all n,

$$\varphi(n) \geq C(\delta)n^\delta.$$

Still equivalent would be to have such a statement for all $n \geq N_0$. This would be very useful provided we actually knew an acceptable value of $C(\delta)$ for some δ, possibly with an explicitly given and reasonably small $N_0(\delta)$ of excluded values. We quote without proof the following convenient result for $\delta = \frac{1}{2}$:

THEOREM 9.5. *For all $n > 6$, $\varphi(n) \geq \sqrt{n}$.*

So in other words, to find all n for which $\varphi(n) \leq 10$, according to this result we need only look at n up to 100, which is fairly reasonable. Of course if you are interested in very large values of φ you will want even stronger bounds. The "truth" is coming up later: there is a remarkable explicit lower bound on $\varphi(n)$.

3. Upper bounds on Euler's φ function

PROPOSITION 9.6. *For any $\epsilon > 0$, there is an n such that $\varphi(n)/n \leq \epsilon$.*

PROOF. Recall that one of our formulas for $\varphi(n)$, or rather for $\varphi(p_1^{a_1} \cdots p_r^{a_r})$, is really a formula for $\varphi(n)/n$:

$$\varphi(n)/n = \prod_{i=1}^{r}(1 - \frac{1}{p_i}).$$

Just for fun, let's flip this over:

$$\frac{n}{\varphi(n)} = \prod_{i=1}^{r}(1 - \frac{1}{p_i});$$

now what we need to show is that for any $L > 0$, we can choose primes $p_1, \ldots, p_r$ such that $\prod_{i=1}^{r}(\frac{p_i-1}{p_i})^{-1} > L$.

Well, at the moment we (sadly for us) don't know much more about the sequence of primes except that it is infinite, so why don't we just take n to be the product of the first r primes $p_1 = 2, \ldots, p_r$? And time for a dirty trick: for any i, $1 \leq i \leq r$, we can view $\frac{1}{1-\frac{1}{p_i}}$ as the sum of a geometric series with ratio $r = \frac{1}{p_i}$. This gives

$$\frac{n}{\varphi(n)} = \prod_{i=1}^{r}(1 - \frac{1}{p_i})^{-1} = \prod_{i=1}^{r}(1 + p_i^{-1} + p_i^{-2} + \ldots).$$

The point here is that if we formally extended this product over all primes:

$$\prod_{i=1}^{\infty}(1 + p_i^{-1} + p_i^{-2} + p_i^{-3} + \ldots)$$

and multiplied it all out, what would we get? A moment's reflection reveals a beautiful surprise: the uniqueness of the prime power factorization is precisely

equivalent to the statement that multiplying out this infinite product we get the infinite series $\sum_{n=1}^{\infty} \frac{1}{n}$, i.e., the harmonic series! Well, except that the harmonic series is divergent. That's actually a good thing; but first let's just realize that if we multiply out the finite product $\prod_{i=}^{r}(1 - \frac{1}{p_i})^{-1}$ we get exactly the sum of the reciprocals of the integers n which are divisible only by the first r primes. In particular – since of course $p_r \geq r$, this sum contains the reciprocal of the first r integers, so: with $n = p_1 \cdots p_r$,

$$\frac{n}{\varphi(n)} \geq \sum_{n=1}^{r} \frac{1}{n}.$$

But now we're done, since as we said before the harmonic series diverges – recall that a very good approximation to the rth partial sum is $\log r$, and certainly $\lim_{r\to\infty} \log r = \infty$. This proves the result. □

To summarize, if we want to make $\varphi(n)/n$ arbitrarily small, we can do so by taking n to be divisible by sufficiently many primes. On the other hand $\varphi(n)/n$ doesn't have to be small: $\varphi(p)/p = \frac{p-1}{p} = 1 - \frac{1}{p}$, and of course this quantity approaches 1 as $p \to \infty$. Thus the relative size of $\varphi(n)$ compared to n depends quite a lot on the shape of the prime power factorization of n.

Contemplation of this proof shows that we had to take n to be pretty darned large in order for $\varphi(n)$ to be significantly smaller than n. In fact this is not far from the truth.

4. The Truth About Euler's φ Function

It is the following:

THEOREM 9.7. *a) For any $\epsilon > 0$ and all sufficiently large n, one has*

$$\frac{\varphi(n) \log\log n}{n} \geq e^{-\gamma} - \epsilon.$$

b) There exists a sequence of distinct positive integers n_k such that

$$\lim_{k\to\infty} \frac{\varphi(n_k) \log\log n_k}{n_j} = e^{-\gamma}.$$

Comments: (a) Here γ is our friend the Euler-Mascheroni constant, i.e.,

$$\lim_{n\to\infty} \sum_{k=1}^{n} (\frac{1}{k}) - \log n \approx 0.5772.$$

(b) What the result is really saying is that $n/\varphi(n)$ can be, for arbitrarily large n, as large as a constant times $\log\log n$, but no larger.

In stating the result in two parts we have just spelled out a fundamental concept from real analysis (which however is notoriously difficult for beginning students to understand): namely, if for any function $f : \mathbb{Z}^+ \to \mathbb{R}$ we have a number L with the property: for every $\epsilon > 0$, then
(i) for all sufficiently large n one has

$$f(n) > L - \epsilon,$$

and (ii) for all $L' < L$ there are only finitely many n such that $f(n) < L'$, then one says that L is the **lower limit** (or limit inferior) of $f(n)$, written

$$\lim \inf_{n\to\infty} f(n) = L.$$

There is a similar definition of the **upper limit** (or limit superior) of a function: it is the largest L such that for any $\epsilon > 0$, for all but finitely many n we have $f(n) < L + \epsilon$. A function which is unbounded below (i.e., takes arbitrarily small values) has no lower limit according to our definition, so instead one generally says that $\liminf f = -\infty$, and similarly we put $\limsup f = +\infty$ when f is unbounded above. With these provisos, the merit of the upper and lower limits is that they *always* exist; moreover one has

$$\liminf f \leq \limsup f$$

always, and equality occurs iff $\lim_{n\to\infty} f$ exists (or is $\pm\infty$). Using this terminology we can summarize the previous results much more crisply:

Since $\varphi(p) = p - 1$, we certainly have

$$\limsup \varphi(n)/n = 1,$$

so we are only interested in how small $\varphi(n)$ can be for large n. We first showed that $\lim_{n\to\infty} \varphi(n) = +\infty$, and indeed that for any $\delta < 1$,

$$\lim_{n\to\infty} \varphi(n)/n^\delta = \infty.$$

However, for $\delta = 1$,

$$\lim \inf_{n\to\infty} \varphi(n)/n = 0.$$

Thus the "lower order" of $\varphi(n)$ lies somewhere between n^δ for $\delta < 1$ (i.e., φ is larger than this for all sufficiently large n) and n (i.e., φ is smaller than this for infinitely many n). In general, one might say that an arithmetic function f has **lower order** $g : \mathbb{Z}^+ \to (0, \infty)$ (where g is presumably some relatively simple function) if

$$\liminf_{n\to\infty} \frac{f}{g} = 1.$$

So the truth is that the lower order of φ is $\frac{e^\gamma n}{\log\log n}$. We will not prove this here.

Remark: all statements about limits, lim inf's lim sup's and so on of a function f, by their nature are independent of the behavior of f on any fixed finite set of values: if we took any arithmetic function and defined it completely randomly for the first $10^{10^{10}}$ values, then we would not change its lower/upper order. However in practice we would like inequalities which are true for all values of the function, or at least are true for an explicitly excluded and reasonably small finite set of values. In the jargon of the subject one describes the latter, better, sort of estimate as an **effective bound**. You can always ask the question "Is it effective?" at the end of any analytic number theory talk and the speaker will either get very happy or very defensive according to the answer. So here we can ask if there is an effective lower bound for φ of the right order of magnitude, and the answer is a resounding yes. Here is a nuclear-powered lower bound for the totient function:

THEOREM 9.8. *For all $n > 2$ we have*

$$\varphi(n) > \frac{n}{e^\gamma \log\log n + \frac{3}{\log\log n}}.$$

PROOF. See [**RS62**, Thm. 15]. □

5. Other Functions

5.1. The sum of divisors function σ. The story for the function σ is quite similar to that of φ. In fact there is a very close relationship between the size of σ and the size of φ coming from the following beautiful double inequality.

PROPOSITION 9.9. *For all n, we have*

$$\frac{1}{\zeta(2)} < \frac{\sigma(n)\varphi(n)}{n^2} < 1.$$

PROOF. Indeed, if $n = \prod_i p_i^{a_i}$, then

$$\sigma(n) = \prod_i \frac{p_i^{a_i+1} - 1}{p_i - 1} = n \prod_i \frac{1 - p^{-a_i-1}}{1 - p_i^{-1}},$$

whereas

$$\varphi(n) = n \prod_i (1 - p_i^{-1}),$$

so

$$\frac{\sigma(n)\varphi(n)}{n^2} = \prod_i (1 - p_i^{-a_i-1}).$$

We have a product of terms in which each factor is less than one; therefore the product is at most 1. Conversely, each of the exponents is less than or equal to -2, so the product is at least as large as the product $\prod_p (1 - p^{-2})$. Now in general, for $s > 1$ we have

$$\prod_p (1 - p^{-s})^{-1} = \prod_p (1 + p^{-s} + p^{-2s} + \ldots) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \zeta(s),$$

so the last product is equal to $\frac{1}{\zeta(2)}$. □

Remark: Recall that $\zeta(2) = \frac{\pi^2}{6}$, so that $\frac{1}{\zeta(2)} = \frac{6}{\pi^2}$.

From this result and the corresponding results for φ we immediately deduce:

THEOREM 9.10. *For every $\delta > 0$, $\frac{\sigma(n)}{n^{1+\delta}} \to 0$.*

In fact we can prove this directly, the same way as for the φ function.

The "truth" about the lower order of φ dualizes to give the true upper order of σ, up to an ambiguity in the multiplicative constant, which will be somewhere between $\zeta(2)^{-1}e^{-\gamma}$ and $e^{-\gamma}$. In fact the latter is correct:

THEOREM 9.11.

$$\limsup_{n\to\infty} \frac{e^{-\gamma}\sigma(n)}{n \log\log n} = 1.$$

And again, because $\sigma(p) = p + 1 \sim p$ for primes, we find that the lower order of $\sigma(n)$ is just n.

5.2. The divisor function. The divisor function $d(n)$ is yet more irregularly behaved than φ and σ, as is clear because $d(p) = 2$ for all primes 2, but of course d takes on arbitrarily large values. In particular the lower order of d is just the constant function 2. As regards the upper order, we limit ourselves to the following two estimates, which you are asked to establish in the homework:

THEOREM 9.12. *For any $\delta > 0$, $\lim_{n\to\infty} \frac{d(n)}{n^\delta} = 0$.*

In other words, for large n, the number of divisors of n is less than any prearranged power of n. This makes us wonder whether its upper order is logarithmic or smaller, but in fact this is not the case either.

PROPOSITION 9.13. *For any $k \in \mathbb{Z}^+$ and any real number C, there exists an n such that $d(n) > C(\log n)^k$.*

Thus the upper order of $d(n)$ is something greater than logarithmic and something less than any power function. We leave the matter there, although much more could be said.

6. Average orders

If we take the perspective that we are interested in the distribution of values of an arithmetic function like φ (or d or σ or ...) in a statistical sense, then we ought to worry that just knowing the upper and lower orders is telling us a very small piece of the story. As a very rough comparison, suppose we tried to study the American educational system by looking at its best and worst students, not in the sense of best or worst ever, but were concerned with what sort of education the top 0.1% and the bottom 0.1% of Americans get, durably over time. The first task is certainly of some interest – for instance, we all wonder how our upper echelon compares to the creme de la creme of the educational systems of other nations and other times; are we producing more or less scientific geniuses and so forth – and the latter task is profoundly depressing, but probably no one will be deluded into believing that we are studying anything like what the "typical" or "average" American learns.

It may interest you to know that the rich range of statistical techniques that can be so fruitfully applied to studying distributions of real-world populations can be equally well applied to study the distribution of values of arithmetic functions. Indeed, this is a flourishing subbranch of analytic number theory: **statistical number theory**. Here we have time only to sample some of the main developments of analytic number theory. And, what is yet more sad, we cannot assume that we know enough about these statistical tools to apply them in all their glory.[1] But probably we are all familiar with the notion of an *average*.

The idea here is that if $f(n)$ is irregularly behaved, we can "smooth it out" by considering its successive averages, say

$$f_a(n) = \frac{1}{n}\sum_{k=1}^{n} f(k).$$

[1]This is one case where by "we" I mean "me"; maybe your knowledge of statistics is equal to the task, but I must confess that mine is not.

We have every right to expect for f_a to be better behaved than f itself, and we say that the **average order** of f is some (presumably simpler) function g if $f_a \sim g$.

As a sample we give the following classic result:

THEOREM 9.14. *The average order of the totient function is* $g(n) = \frac{1}{2\zeta(2)}n = \frac{3}{\pi^2}n$.

Thus "in some sense" the typical value of $\varphi(n)/n$ is about .304. It would be nice to interpret this as saying that if we pick an n at random, then with large probability $\varphi(n)/n$ is close to .304, but of course we well know that the average – i.e., the arithmetic mean – does not work that way. Just because the average score on an exam is 78 does not mean that most students got a grade close to 78. (Perhaps 2/3 of the course got A grades and the other third failed; these things do happen.) Nevertheless it is an interesting result, and to prove it we will derive a very interesting consequence.

First however it is nice to have a "harder analysis" analogue of Theorem 9.14. That is, the theorem at the moment asserts that $\frac{1}{n}\sum_{k=1}^n \varphi(k) \sim \frac{3}{\pi^2}n$, or equivalently

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \varphi(k)}{\frac{3}{\pi^2}n^2} = 1.$$

This in turn means that if we define "the error term" $E(n) = \sum_{k=1}^n \varphi(k) - (\frac{3}{\pi^2})n^2$, so that $\sum_{k=1}^n \varphi(k) = \frac{3}{\pi^2}n^2 + E(n)$, then the error term is small compared to the main term: namely it is equivalent to

$$\lim_{n\to\infty} \frac{E(n)}{(3/\pi^2)n^2} = 0.$$

So far we are just pushing around the definitions. But a fundamentally better thing to do would be to give an upper bound on the error $E(n)$, i.e., to find a nice, simple function $h(n)$ such that $E(n) \leq h(n)$ and $\frac{h(n)}{(3/\pi^2)n^2} \to 0$ In fact we do not need $E(n) \leq h(n)$ quite: if $E(n)$ were less than $100h(n)$ that would be just as good, because if $h(n)$ divided by $(3/\pi^2)n^2$ approaches zero, the same would hold for $100h(n)$. This motivates the following notation:

We say that $f(n) = O(g(n))$ if there exists a constant C such that for all n, $f(n) \leq Cg(n)$. So it would be enough to show that $E(n) = O(h(n))$ for some function h which approaches zero when divided by $\frac{3}{\pi^2}n^2$, or in more colloquial language, for any function grows less than quadratically. So for instance a stronger statement would be that $E(n) = O(n^\delta)$ for any $\delta < 2$. In fact one can do a bit better than this:

THEOREM 9.15.

$$\sum_{k=1}^n \varphi(k) = \frac{3}{\pi^2}n^2 + O(n \log n).$$

Or again, dividing through by n, this theorem asserts that that the average over the first n values of the ϕ function is very nearly $\frac{3}{\pi^2}n$, and the precise sense in which this is true is that the difference between the two is bounded by a constant times $\log n$. Note that it would be best of all to know an actual acceptable value of such

a constant – that would be a "completely effective" version of the statement, but we do not go into this here (or need to).

Now we can give a very nice application, which we can first state in geometric language. It concerns the lattice points in the plane, i.e., the numbers $(x, y) \in \mathbb{R}^2$ with both x and y integers (so, if you like, the Gaussian integers, although this extra structure is not relevant here). Suppose we are some lord of the lattice points, sitting at the origin $(0, 0)$ and surveying our entire domain. What do we see? Well – assuming a truly 2-dimensional situation – we can see some of our loyal subjects but not others. For instance we can certainly see the point $(1, 1)$, but we cannot see $(2, 2)$, $(3, 3)$ or any $n(1, 1)$ with $n > 1$ since the point $(1, 1)$ itself obscures our view.

Thus we can define a lattice point (x, y) to be **visible** (from the origin) if the line segment from $(0, 0)$ to (x, y) contains no lattice points on its interior. Suppose we start coloring the lattice, coloring a lattice point red if we can see it and black if we cannot see it. Try it yourself: this gives a very interesting pattern. It is natural to ask: how many of the lattice points can we see?

Well, the first observation is that a lattice point (x, y) is visible iff $\gcd(x, y) = 1$: an obstructed view comes precisely from a nontrivial common divisor of x and y. From this it follows that the answer is "infinitely many": for instance we can see $(1, n)$ for all integers n, and many more besides. Well, let us change our question a bit. Suppose that each of the lattice points is supposed to pay a flat tax to our lordship, and if we see a lattice point then we can see whether or not it has paid its taxes. What percentage of our revenue are we collecting if we only worry about the lattice points we can see?

Now to formalize the question. If we ask about the entire lattice at once, the answer to most of our questions is always going to be "infinity," and moreover an actual king (even a two-dimensional one) probably rules over a finite kingdom. So for a positive integer N, let us write $L(N)$ for the number of lattice points (x, y) with $|x|,\ |y| \leq N$ – that is, the lattice points lying in the square centered at the origin with length (and width) equal to $2N$. Well, this number is $(2N + 1)^2$: there are $2N + 1$ possible values for both x and y. But now define $V(N)$ to be the number of visible lattice points, and our question is: when N is large, what can we say about $\frac{V(N)}{L(N)}$?

THEOREM 9.16. *We have* $\lim_{N\to\infty} \frac{V(N)}{L(N)} = \frac{6}{\pi^2}$.

Before we prove the result, we can state it in a slightly different but equally striking way. We are asking after all for the number of ordered pairs of integers (x, y) each of absolute value at most N, with x and y relatively prime. So, with a bit of poetic license perhaps, we are asking: what is the probability that two randomly chosen integers are relatively prime? If we lay down the ground rules that we are randomly choosing x and y among all integers of size at most N, then the astonishing answer is that we can make the probability as close to $\frac{6}{\pi^2}$ as we wish by taking N sufficiently large.

Now let us prove the result, or at any rate deduce it from Theorem 9.14. First we observe that the eight lattice points immediately nearest the origin – i.e., those with $\max(|x|,|y|) \leq 1$ – are all visible. Indeed there is an eightfold symmetry in the situation: the total number of visible lattice points in the square $|x|,|y| \leq N$ will then be these 8 plus 8 times the number of lattice points with $2 \leq x \leq N$, $1 \leq y \leq x$ (i.e., the ones whose angular coordinate θ satisfy $0 < \theta \leq \frac{\pi}{2}$. But now we have

$$V(N) = 8 + \sum_{2 \leq n \leq N} \sum_{1 \leq m \leq n, (m,n)=1} 1 = 8 \sum_{1 \leq n \leq N} \varphi(N).$$

Aha: we know that $\sum_{1 \leq n \leq N} \varphi(N) = \frac{3}{\pi^2} N^2 + O(N \log N)$, so

$$|\frac{V(N) - \frac{24}{\pi^2} N^2}{L(N)}| \leq C \frac{N \log N}{L(N)}.$$

But now $L(N) = (2N+1)^2$, and $C \frac{N \log N}{N^2} \to 0$ as $N \to \infty$, so we find that

$$0 = \lim_{N \to \infty} \frac{V(N) - \frac{24}{\pi^2} N^2}{(2N+1)^2},$$

or

$$\lim_{N \to \infty} \frac{V(N)}{L(N)} = \lim_{N \to \infty} \frac{\frac{24}{\pi^2}}{(2 + \frac{1}{N})^2} = \frac{6}{\pi^2}.$$

Having given a formal proof of this result based upon the unproved Theorem 9.14 (trust me that this theorem is not especially difficult to prove; it just requires a bit more patience than we have at the moment), let us now give a proof which is not rigorous but is extremely interesting and enlightening. Namely, what does it "really mean" for two integers x and y to be relatively prime? It means that there is no prime number p which simultaneously divides both x and y. Remarkably, this observation leads directly to the result. Namely, the chance that x is divisible by a prime p is evidently $\frac{1}{p}$, so the chance that x and y are both divisible by p is $\frac{1}{p^2}$. Therefore the chance that x and y are not both divisible by a prime p is $(1 - p^{-2})$. Now we think of being divisible by different primes as being "independent" events: if I tell you that an integer is divisible by 3, not divisible by 5 and divisible by 7, and ask you what are the odds it's divisible by 11, then we still think the chance is $\frac{1}{11}$. Now the probability that each of a set of independent events all occur is the product of the probabilities that each of them occur, so the probability that x and y are not simultaneously divisible by any prime p ought to be $(1-2^{-2}) \cdot (1-3^{-2} \cdot \ldots (1-p^{-2}) \cdots$, but we saw earlier that this infinite product is nothing else than the reciprocal of $\sum_{n=1}^{\infty} \frac{1}{n^2} = \zeta(2)$. Thus the answer should be $\frac{1}{\zeta(2)} = \frac{6}{\pi^2}$!

The argument was not rigorous because the integers are not really a probability space: there is nothing random about whether, say, 4509814091046 is divisible by 103; either it is or it isn't. Instead of probability one should rather work with the notion of the "density" of a set of integers (or of a set of pairs of integers) – a notion which we shall introduce rather soon – and then all is well until we pass from sets defined by divisibility conditions on finitely many primes to divisibility conditions on all (infinitely many) primes. This is not to say that such a "probability-inspired proof" cannot be pushed through – it absolutely can. Moreover, the fact that the

probabilistic argument gives an answer which can be proven to be the correct answer via conventional means is perhaps most interesting of all.

Finally, we note that probabilistic reasoning gives the same answer to a closely related question: what is the probability that a large positive integer n is squarefree? This time we want, for each prime p "independently", n not to be divisible by p^2, of which $1 - p^{-2}$ percent of all integers are. Therefore we predict that the probability that n is squarefree is also $\frac{6}{\pi^2}$, and this too can be proved by similar (although not identical) means to the proof of Theorem 9.14.

6.1. The average order of the Möbius function. We are interested in the behavior of

$$\mu_a(n) = \frac{1}{n} \sum_{k=1}^{n} \mu(k).$$

This is a horse of a completely different color, as we are summing up the values 0 and ± 1. We just saw that μ is nonzero a positive proportion, namely $\frac{6}{\pi}$, of the time. Looking at values of the Möbius function on squarefree integers one finds that it is indeed $+1$ about as often as it is -1, which means that there ought to be a lot of cancellation in the sum. If every single term in the sum were 1 then $\mu_a(n)$ would still only be equal to 1, and similarly if every single term were -1 the average order would be -1, so the answer (if the limit exists!) is clearly somewhere in between. The only sensible guess turns out to be true:

THEOREM 9.17. *The average order of the Möbius function is zero:*

$$\lim_{n \to \infty} \frac{\sum_{k=1}^{n} \mu(k)}{n} = 0.$$

However, this turns out to be a formidably deep result. Namely, it has been known for almost two hundred years that this statement is logically equivalent to the single greatest result in analytic number theory: the prime number theorem (coming up!). However the prime number theorem has only been known to be true for a bit over one hundred years.

In fact one can ask for more: the statement that the average order of μ is zero is equivalent to the statement that the sum $\sum_{k=1}^{n} \mu(k)$ is of smaller order than n itself, i.e., we need just enough cancellation to beat the trivial bound. But if you do some computations you will see that these partial sums seem in practice to be *quite a bit* smaller than n, and to say exactly how large they should be turns out to be a much deeper problem yet: it is equivalent to the **Riemann hypothesis**. I hope to return to this later, but for now I leave you with the question: suppose you believed that the nonzero values of the Möbius function were *truly* random: i.e., for every n we flip a coin and bet on heads: if we win, we add 1 to the sum, and if we lose, we subtract 1 from the sum. It is then clearly ridiculous to expect to win or lose all the games or anything close to this: after n games we should indeed be much closer to even than to having won or lost n dollars. But how close to even should we expect to be?

CHAPTER 10

The Primes: Infinitude, Density and Substance

1. The Infinitude of the Primes

The title of this section is surely, along with the uniqueness of factorization, the most basic and important fact in number theory. The first recorded proof was by Euclid, and we gave it at the beginning of the course. There have since been (very!) many other proofs, many of which have their own merits and drawbacks. It is entirely natural to look for further proofs: in terms of the arithmetical function $\pi(n)$ which counts the number of primes $p \leq n$, Euclid's proof gives that

$$\lim_{n\to\infty} \pi(n) = \infty.$$

After the previous section we well know that one can ask for more, namely for the asymptotic behavior (if any) of $\pi(n)$. The asymptotic behavior is known – the celebrated **Prime Number Theorem**, coming up soon – but it admits no proof simple enough to be included in this course. So it is of interest to see what kind of bounds (if any!) we get from some of the proofs of the infinitude of primes we shall discuss.

1.1. Euclid's proof. We recall Euclid's proof. There is at least one prime, namely $p_1 = 2$, and if $p_1, \ldots, p_n$ are any n primes, then consider

$$N_n = p_1 \cdots p_n + 1.$$

This number N_n may or may not be prime, but being at least 3 it is divisible by some prime number q, and we cannot have $q = p_i$ for any i: if so $p_i | p_1 \cdots p_n$ and $p_i | N_n$ implies $p_i | 1$. Thus q is a new prime, which means that given any set of n distinct primes we can always find a new prime not in our set: therefore there are infinitely many primes.

Comments: (i) Euclid's proof is often said to be "indirect" or "by contradiction", but this is unwarranted: given any finite set of primes $p_1, \ldots, p_n$, it gives a perfectly definite procedure for constructing a new prime.

(ii) Indeed, if we define $E_1 = 2$, and having defined $E_1, \ldots, E_n$, we define E_{n+1} to be the smallest prime divisor of $E_1 \cdots E_n + 1$, we get a sequence of distinct prime numbers, nowadays called the **Euclid sequence** (of course we could get a different sequence by taking p_1 to be a prime different from 2). The Euclid sequence begins

$$2, 3, 7, 43, 13, 53, 5, \ldots$$

Many more terms can be found on the *online handbook of integer sequences*. The obvious question – does every prime occur eventually in the Euclid sequence with $p_1 = 2$ (or in any Euclid sequence?) remains unanswered.

(iii) It is certainly a "classic" proof, but it is not "aesthetically perfect" (whatever that may mean). Namely, there is a moment when the reader wonders – hey, why are we multiplying together the known primes and adding one? One can address this by pointing out in advance the key fact that $\gcd(n, n+1) = 1$ for all n. Therefore if there were only finitely many primes $p_1, \ldots, p_r$, there would be an integer divisible by all of them, $N = p_1 \cdots p_r$, and then the fact that $\gcd(N, N+1) = 1$ leads to a contradiction. I do like this latter version better, but it is really just a rewording of Euclid's proof.[1]

(iv) Euclid's proof can be used to prove some further results. For instance:

THEOREM 10.1. *Fix a positive integer $N > 2$. Then there are infinitely many primes p which are* not *congruent to* $1 \pmod N$.

PROOF. Take $p_1 = 2$, which is not congruent to $1 \pmod N$. Assume that $p_1, \ldots, p_n$ is a list of n primes, none of which are $1 \pmod N$. Now consider the product

$$P_n := Np_1 \cdot \ldots \cdot p_n - 1.$$

$P_n \geq N - 1 \geq 2$, so it has a prime divisor. Also $P_n \equiv -1 \pmod N$. So if every prime divisor q of P_n were 1 mod N, then so would P_n be $1 \pmod N$ – which it isn't – therefore P_n has at least one prime divisor q which is not $1 \pmod N$. As above, clearly $q \neq p_i$ for any i, which completes the proof. □

In fact this argument can be adapted to prove the following generalization.

THEOREM 10.2. *Fix a positive integer $N > 2$, and let H be a proper subgroup of $U(N) = (\mathbb{Z}/N\mathbb{Z})^\times$. There are infinitely many primes p such that $p \pmod N \notin H$.*

The proof is left as an exercise. (Suggestion: fix $a \in \mathbb{Z}^+$, $1 < a < N$, such that $a \pmod N \notin H$. Take $P_0 = 2N + a$ and for $n \geq 1$, $P_n = (2N \prod_{i=1}^n p_i) + a$.)

Remark: If $\varphi(N) = 2$ – that is, for $N = 3$, 4, or 6 – then ± 1 gives a reduced residue system modulo N, so that any prime $p > N - 1$ which is not $1 \pmod N$ is necessarily $-1 \pmod N$. Thus the argument shows that there are infinitely many primes p which are $-1 \pmod 3$, $-1 \pmod 4$ or $-1 \pmod 6$.

Remark: Interestingly, one can also prove without too much trouble that there are infinitely many primes $p \equiv 1 \pmod N$: the proof uses cyclotomic polynomials.

THEOREM 10.3. *For any field F, there are infinitely many irreducible polynomials over F, i.e., infinitely many irreducible elements in $F[T]$.*

PROOF. Euclid's argument works here: take e.g. $p_1(t) = t$, and having produced $p_1(t), \ldots, p_r(t)$, consider the irreducible factors of $p_1(t) \cdots p_r(t) + 1$. □

Note that one can even conclude that there are infinitely many prime ideals in $F[t]$ – equivalently, there are infinitely many *monic* irreducible polynomials. When F is infinite, the monic polynomials $t - a$ for $a \in F$ do the trick. When F is

[1] I have heard this argument attributed to the great 19th century algebraist E. Kummer. For what little it's worth, I believe I came up with it myself as an undergraduate. Surely many others have had similar thoughts.

finite, we showed there are infinitely many irreducible polynomials, but there are only $\#F - 1$ different leading coefficients, so there must be infinitely many monic irreducible polynomials. It is interesting to think about why this argument does not work in an arbitrary PID.[2]

1.2. Fermat numbers. Another way to construe Euclid's proof is that it suffices to find an infinite sequence n_i of pairwise coprime positive integers, because these integers must be divisible by different primes. The Euclid sequence is such a sequence. A more "natural" looking sequence is the following.

THEOREM 10.4. *The Fermat numbers $F_n = 2^{2^n} + 1$ are pairwise coprime.*

PROOF. We claim that for all $n \geq 1$ we have

$$F_n = \prod_{d=0}^{n-1} F_d + 2.$$

This certainly suffices, since if p is some common prime divisor of F_d (for any $d < n$) and F_n then $p \mid F_n - 2$, hence $p \mid 2$, but all the Fermat numbers are odd. The claim itself can be established by induction; we leave it to the reader. □

1.3. Mersenne numbers. Recall that Fermat believed that all the Fermat numbers were prime, and this is not true, since e.g.

$$F_5 = 2^{2^5} + 1 = 641 \cdot 6700417,$$

and in fact there are no larger known prime Fermat numbers. Nevertheless the previous proof shows that there is something to Fermat's idea: namely, they are "almost" prime in the sense that no two of them have a common divisor. One then wonders whether one can devise a proof of the infinitude of the primes using the Mersenne numbers $2^p - 1$, despite the fact that it is unknown whether there are infinitely many Mersenne primes. This can indeed be done:

Let p be a prime (e.g. $p = 2$, as usual) and q a prime divisor of $2^p - 1$. Then $2^p \equiv 1 \pmod{q}$. In other words, p is a multiple of the order of 2 in the cyclic group $(\mathbb{Z}/q\mathbb{Z})^\times$. Since p is prime the order of 2 must be exactly p. But by Lagrange's theorem, the order of an element divides the order of the group, which is $\varphi(q) = q - 1$, so $p \mid q - 1$ and hence $p < q$. Thus we have produced a prime larger than the one we started with.

1.4. Euler's first proof. It is a remarkable fact that the formal identity

$$\prod_p (1 - \frac{1}{p})^{-1} = \sum_n \frac{1}{n}$$

– which amounts to unique factorization – immediately implies the infinitude of primes. Indeed, on the left hand side we have a possibly infinite product, and on the right-hand side we have an infinite sum. But the infinite sum is well-known to be divergent, hence the product must be divergent as well, but if it were a finite product it would certainly be convergent!

[2]For there are PID's with only finitely many prime ideals: e.g. the set of rational numbers whose reduced denominator is prime to 42 is a PID with exactly three prime ideals.

Many times in the course we have seen a rather unassuming bit of abstract algebra turned into a mighty number-theoretic weapon. This example shows that the same can be true of analysis.

1.5. Chaitin's proof. In his most recent book[3] the computer scientist Gregory Chaitin announces an "algorithmic information theory" proof of the infinitude of primes. He says: if there were only finitely many primes $p_1, \ldots, p_k$ then every positive integer N could be written as

$$N = p_1^{a_1} \cdots p_k^{a_k},$$

which is "too efficient" a way of representing all large integers N. Chaitin compares his proof with Euclid's proof and Euler's proof (with a grandiosity that I confess I find unjustified and unbecoming). But criticism is cheaper than understanding: can we at least make sense of his argument?

Let us try to estimate how many integers n, $1 \leq n \leq N$, could possibly be expressed in the form $p_1^{a_1} \cdots p_k^{a_k}$, i.e., as powers of a fixed set of k primes. In order for this expression to be at most N, every exponent has to be much smaller than N: precisely we need $0 \leq a_i \leq \log_{p_i} N$; the latter quantity is at most $\log_2 N$, so there are at most $\log_2 N + 1$ choices for each exponent, or $(\log_2 N + 1)^k$ choices overall. But aha – this latter quantity is much smaller than N when N is itself large: it is indeed the case that the percentage of integers up to N which we can express as a product of any k primes tends to 0 as N approaches infinity.

So Chaitin's proof is indeed correct and has a certain admirable directness to it.

1.6. Another important proof. However, the novelty of Chaitin's proof is less clear. Indeed, in many standard texts (including [**HW**], which was first written in 1938), one finds the following argument, which is really a more sophisticated version of Chaitin's proof.

Again, we will fix k and estimate the number of integers $1 \leq n \leq N$ which are divisible only by the first k primes $p_1, \ldots, p_k$, but this time we use a clever trick: recall that n can be written uniquely as uv^2 where u is squarefree. The number of squarefree u's – however large! – which are divisible only by the first k primes is 2^k (for each p_i, we either choose to include it or not). On the other hand, $n = uv^2 \leq N$ implies that $v^2 \leq N$ and hence $v \leq N^{\frac{1}{2}}$. Hence the number of $n \leq N$ divisible only by the first k primes is at most $2^k\sqrt{N}$. If there are k primes less than or equal to N, we therefore have

$$2^k\sqrt{N} \geq N$$

or

$$k \geq \frac{\log_2(N)}{2}.$$

[3]Its title is *Meta Math! The Quest for Omega*.

1.7. An algebraic number theory proof. We now sketch a proof due to Lawrence Washington.

Let R be a PID, with field of fractions F. Suppose K is a finite-degree field extension of F – in other words, there exists some positive integer d and elements $x_1, \ldots, x_d$ of K such that every element x of K can be written as $\alpha_1 x_1 + \ldots + \alpha_d x_d$ for $\alpha_i \in F$. In such a situation we can define a subset R_L of L, which is the set of all elements x of L which satisfy a monic polynomial relation with coefficients in R: that is, for some $n \in \mathbb{Z}^+$,

$$x^n + r_{n-1}x^{n-1} + \ldots + r_1 x + r_0 = 0,$$

and $r_i \in R$ for all i. It can be shown that R_L is a subring of L, called the **integral closure** of R in L. As an example, when $R = \mathbb{Z}$ and D is a squarefree integer not congruent to 1 (mod 4), then taking $K = \mathbb{Q}$ and $L = \mathbb{Q}(\sqrt{D})$, then the integral closure of $\mathbb{Z}$ in L is our friend the quadratic ring $\mathbb{Z}[\sqrt{D}]$. Anyway, here is the result:

THEOREM 10.5. *Suppose that a PID R, with quotient field K, has only finitely many prime ideals. Then for any finite-degree field extension L of K, the integral closure S of R in L is again a PID.*

This shows the infinitude of the primes in $\mathbb{Z}$, since we saw that $\mathbb{Z}[\sqrt{-5}]$ is *not* a PID!

The proof of Theorem 10.5 lies further up and further in the realm of algebraic number theory than we dare to tread in this course. But here is a sketch of a proof for the "slummers"[4]: the ring S is a Dedekind domain, so for any nonzero prime ideal $\mathfrak{p}$ of R, $\mathfrak{p}S$ is a nontrivial finite product of powers of prime ideals. The distinct prime ideals $\mathcal{P}_i$ appearing in this factorization are precisely the prime ideals $\mathcal{P}$ *lying over* $\mathfrak{p}$, i.e., such that $\mathcal{P} \cap R = \mathfrak{p}$. This shows that the restriction map $\mathcal{P} \mapsto \mathcal{P} \cap R$ from prime ideals of S to prime ideals of R has finite fibers. Thus, since by assumption there are only finitely many prime ideals of R, there are only finitely many prime ideals of S. Finally, a Dedekind domain with only finitely many prime ideals is necessarily a PID, as can be shown using the Chinese Remainder Theorem.

This is a proof with a moral: we need to have infinitely many primes in order for number theory to be as complicated as it is.

1.8. Furstenberg's proof. The last proof we will give is perhaps the most remarkable one. In the 1955 issue of the American Mathematical Monthly there appeared the following article by Hillel Furstenberg, which we quote in its entirety:

"In this note we would like to offer an elementary 'topological' proof of the infinitude of the prime numbers. We introduce a topology into the space of integers S, by using the arithmetic progressions (from $-\infty$ to $+\infty$) as a basis. It is not difficult to verify that this actually yields a topological space. In fact under this topology S may be shown to be normal and hence metrizable. Each arithmetic progression is closed as well as open, since its complement is the union of other arithmetic progressions (having the same difference). As a result the union of any finite number of arithmetic progressions is closed. Consider now the set $A = \cup A_p$, where A_p consists of all multiples of p, and p runs though the set of primes ≥ 2.

[4] i.e., more advanced readers who are reading these notes

The only numbers not belonging to A are -1 and 1, and since the set $\{-1, 1\}$ is clearly not an open set, A cannot be closed. Hence A is not a finite union of closed sets which proves that there are an infinity of primes."

Remarks: Furstenberg was born in 1935, so this ranks as one of the leading instances of undergraduate mathematics in the 20th century. He is now one of the leading mathematicians of our day. What is all the more remarkable is that this little argument serves as a preview of the rest of his mathematical career, which has concentrated on applying topological and dynamical methods ("ergodic theory") to the study of problems in number theory and combinatorics.

2. Bounds

Let us now go through some of these proofs and see what further information, if any, they yield on the function $\pi(n)$.

1. From Euclid's proof one can deduce that $\pi(n) \geq C \log\log n$. We omit the argument, especially since the same bound follows more readily from the Fermat numbers proof. Of course this is a horrible bound.

2. The Mersenne numbers proof gives, I believe, an even worse (iterated logarithmic) bound. I leave it to the reader to check this.

3. Euler's first proof does not immediately come with a bound attached to it. However, as we saw earlier in our study of the ϕ function, it really shows that

$$\prod_{i=1}^{r}(1 - \frac{1}{p_i})^{-1} > \sum_{i=1}^{r} \frac{1}{r} \geq C \log r.$$

After some work, one can deduce from this that

$$\sum_{i=1}^{n} \frac{1}{p_i} \geq C \log\log n,$$

whence the divergence of the prime reciprocals. We will not enter into the details.

4. Chatin's proof gives a lower bound on $\pi(n)$ which is between $\log\log n$ and $\log n$ (but much closer to $\log n$).

5. As we saw, one of the merits of the proof of §1.6 is that one easily deduces the bound $\pi(n) \geq \frac{\log_2 n}{2}$. (Of course, this is still almost a full exponential away from the truth.)

6. As we mentioned, knowing that the prime reciprocals diverge *suggests* that $\pi(n)$ is at worst only slightly smaller than n itself. It *shows* that $\pi(n)$ is not bounded above by any power function Cn^{δ} for $\delta < 1$.

7. The last two proofs give no bounds whatsoever, not even implicitly. This seems to make them the worst, but there are situations in which one wants to separate out the problem of proving the infinitude of a set of numbers from the problem of estimating its size, the latter problem being either not of interest or (more often)

hopelessly out of current reach. In some sense all of the arguments except the last two are implicitly *trying to prove too much* in that they give lower bounds on $\pi(n)$. Trying to prove more than what you really want is often a very good technique in mathematics, but sometimes, when the problem is really hard, making sure that you are concentrating your efforts solely on the problem at hand is also a key idea. At any rate, there are many problems in analytic combinatorics for which Furstenberg-type existence proofs either were derived long before the explicit lower bounds (which require much more complicated machinery) or are, at present, the only proofs which are known.

3. The Density of the Primes

All of the results of the previous section were *lower bounds* on $\pi(x)$. It is also of interest to give an upper bound on $\pi(x)$ beyond the "trivial" bound $\pi(x) \leq x$. The following gives such a result.

THEOREM 10.6. *As $n \to \infty$, we have $\frac{\pi(n)}{n} \to 0$.*

If you like, this result expresses that the probability that a randomly chosen positive integer is prime is 0. We will come back to this idea after the proof, replacing "probability" by the more precise term *density*.

PROOF. Let us first observe that there are at most $\frac{N}{2}$ primes in the interval $[1, N]$ since all but one of them must be odd. Similarly, since only one prime is divisible by 3, every prime $p > 6$ must be of the form $6k + 1$ or $6k + 5$, i.e., only 2 of the 6 residue classes mod 6 can contain more than one prime (in fact some of them, like 4, cannot contain any primes, but we don't need to worry about this), so that of the integers $n \leq N$, at most $\frac{2}{6}N + 6 + 2$ are primes.

In fact this simple reasoning can be carried much farther, using what we know about the φ function. Namely, for any positive integer d, if $\gcd(a, d) > 1$ there is at most one prime $p \equiv a \pmod{d}$.[5] In other words, only $\varphi(d)$ out of d congruence classes mod d can contain more than one prime, so at most $(\frac{\varphi(d)}{d})N + d + \varphi(d)$ of the integers $1 \leq n \leq N$ can possibly be prime. (Here we are adding d once to take care of the one prime that might exist in each congruence class and adding d a second time to take care of the fact that since N need not be a multiple of d, so the "partial congruence class" at the end may contain a higher frequency of primes than $\varphi(d)/d$, but of course no more than $\varphi(d)$ of primes overall.) But we know, thank goodness, that for every $\epsilon > 0$, there exists a d such that $\frac{\varphi(d)}{d} < \epsilon$, and choosing this d we find that the number of primes $n \leq N$ is at most

$$\frac{\pi(N)}{N} \leq \frac{\epsilon N + d + \varphi(d)}{N} = \epsilon + \frac{d + \varphi(d)}{N}.$$

This approaches ϵ as $N \to \infty$, so is, say, less than 2ϵ for all sufficiently large N. □

Remark: Reflecting on the proof, something slightly strange has happened: we showed that $\varphi(d)/d$ got arbitrarily small by evaluating at $d = p_1 \cdots p_r$, the product of the first r primes. Thus, in order to show that the primes are relatively sparse, we used the fact that there are infinitely many of them!

[5]Recall this is true because if $x \equiv a \pmod{d}$, $\gcd(a, d) \mid d \mid x - a$, and $gcd(a, d) \mid a$, so $\gcd(a, d)|x$.)

In fact, by similarly elementary reasoning, one can prove a more explicit result, that $\pi(n) \leq \frac{Cn}{\log\log n}$. Before moving on to discuss some similar and stronger statements about the order of magnitude of $\pi(n)$, let us digress a bit on the notion of density of a set of integers.

Definition: A subset A of the positive integers is said to have **density** $\delta(A) = \alpha$ if

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N \mid n \in A\}}{N} = \alpha.$$

We have encountered this notion before: recently we claimed (based on something less than a rigorous proof) that the density of squarefree integers is $\frac{6}{\pi^2}$. Notice that we have just shown that the primes have density zero. Here are some further examples:

Example 1: For any positive integers a and N, the density of the set of positive integers $x \equiv a \pmod N$ is $\frac{1}{N}$. In particular, the set of all positive integers whose last decimal digit is 1 has density $\frac{1}{10}$.

Example 2: Any finite set has density zero.

Example 3: For any $k > 1$, the set of kth powers n^k has density 0.

Example 4: In fact the set of all *proper powers*, i.e., positive integers of the form n^k with $k > 1$, has density zero.

Example 5: The set A of all positive integers whose *first* decimal digit is 1 does not have a density: that is, the limit does not exist. To see this, let $C(N)$ be the number of positive integers $1 \leq n \leq N$ with first digit 1. For any $k \geq 1$, $C(2 \cdot 10^k - 1) \geq \frac{1}{2}(2 \cdot 10^k - 1)$, since all of the integers from 10^k to $2 \cdot 10^k - 1$ begin with 1, and this is half of all integers less than $2 \cdot 10^k - 1$. On the other hand, none of the integers from $2 \cdot 10^k$ to $10^{k+1} - 1$ begin with 1, and this is $\frac{8}{10}$ of the integers less than or equal to $10^{k+1} - 1$, so $C(10^{k+1} - 1) \leq \frac{2}{10}(10^{k+1} - 1)$. Thus $C(N)/N$ does not tend to any limiting value.

Because of this definition it is common to discuss also the upper density $\overline{\delta}(A)$ and the lower density $\underline{\delta}(A)$: in the above definition replace lim by lim inf (resp. lim *sup*), the point being that these two quantities exist for any set, and a set A has a density if $\underline{\delta} = \overline{\delta}$. Note that if $\overline{\delta}(A) = 0$, then necessarily $\delta(A)$ exists and equals 0.

Example 6: If $A_1, \ldots, A_k$ are finitely many sets having densities $\alpha_1, \ldots, \alpha_k$, respectively, then the upper density of $A_1 \cup \ldots \cup A_k$ is at most $\alpha_1 + \ldots + \alpha_k$. If $A_1, \ldots, A_k$ are pairwise disjoint, then the density of $A_1 \cup \ldots \cup A_k$ exists and is exactly $\alpha_1 + \ldots + \alpha_k$. (In fact it is enough if the pairwise intersections $A_i \cap A_j$ all have density zero.)

Density versus probability: Why have we backed off from using the word "probability"? Because ever since the work of the great early twentieth century Russian

mathematician Kolmogorov, mathematicians have been trained to use the word "probability" only in the measure-theoretic sense, or, in plainer language, for the following situation: we have a set S (the "sample space") and a function which associates to each reasonable subset E (an "event") a number $P(E)$, $0 \leq P(E) \leq 1$, and satisfying the axiom of **countable additivity**: if $\{E_i\}_{i=1}^{\infty}$ is a sequence of events which are strongly mutually exclusive (i.e., $E_i \cap E_j = \emptyset$ for all $i \neq j$), then

$$P(\bigcup_i E_i) = \sum_{i=1}^{\infty} P(E_i).$$

Our density function δ satisfies finite additivity but not countable additivity: indeed, if we took A_i to be the singleton set $\{i\}$, then certainly $\delta(A_i) = 0$ for all i but the union of all the A_i's are the positive integers themselves, so have density 1. This is the problem: for a probability *measure* we cannot have (countably!) infinitely many sets of measure zero adding up to a set of positive measure, but this happens for densities.

A similar problem occurs in our "proof" that the squarefree integers have density $\frac{6}{\pi^2}$. The set S_{p^2} of integers which are *not* multiples of p^2 has density $1 - \frac{1}{p^2}$, and it is indeed true that these sets are "finitely independent" in the sense that the intersection of any finite number of them has density equal to the product of the densities of the component sets:

$$\delta(S_{p_1} \cap \ldots \cap S_{p_n}) = \prod_{i=1}^{n} (1 - \frac{1}{p^2}).$$

4. Substance

Let us define a subset S of the positive integers to be **substantial** if $\sum_{n \in S} \frac{1}{n} = \infty$.

Example 0: Obviously a finite set is not substantial.

Example 1: The set $\mathbb{Z}^+$ is substantial: the harmonic series diverges.

Example 2: If S and T are two sets with finite symmetric difference – that is, there are only finitely many elements that are in S and not T or in T but not S – then S is substantial iff T is substantial.

Example 3: For any subset S of $\mathbb{Z}^+$, at least one of S and its complementary subset $S' = \mathbb{Z}^+ \setminus S$ is substantial, since

$$\sum_{n \in S} \frac{1}{n} + \sum_{n \in S'} \frac{1}{n} = \sum_{n \in \mathbb{Z}^+} \frac{1}{n} = \infty.$$

So there are "plenty" of substantial subsets. It is certainly possible for both S and S' to be substantial: take, e.g. the set of even numbers (or any S with $0 < \delta(S) < 1$: see below).

Example 4: For any fixed $k > 1$, the set of all perfect kth powers is not substantial: by (e.g.) the Integral Test, $\sum_{n=1}^{\infty} \frac{1}{n^k} < \infty$.

Example 5: The set of integers whose first decimal digit is 1 is substantial.

Example 6: Indeed any set S with positive upper density is substantial. This is elementary but rather tricky to show, and is left as a (harder) exercise.

The converse does not hold. Indeed, we saw above that the primes have zero density, but we will now establish the following:

THEOREM 10.7. *The sum $\sum_p \frac{1}{p}$ of the prime reciprocals is infinite.*

PROOF. (Erdős) Seeking a contradiction, we suppose that the series converges: then there exists an k such that

$$\sum_{p>p_k} \frac{1}{p} < \frac{1}{2}.$$

However, the number of integers $1 \leq n \leq N$ which are divisible by p_{k+1} is at most $\frac{N}{p_{k+1}}$; similarly for p_{k+2}, p_{k+3}, so that overall the number of integers which are divisible by any $p > p_k$ is at most

$$\frac{N}{p_{k+1}} + \frac{N}{p_{k+2}} + \ldots = N \sum_{p>p_k} \frac{1}{p} = \frac{N}{2}.$$

But this says that for any N, at least half of all positive integers are divisible by one of the first k primes, which the argument of §1.6 showed not to be the case. □

Remarks: Maybe this is the best "elementary" proof of the infinitude of the primes. Aside from being an elegant and interesting argument, it is a quantum leap beyond the previous results: since for any $k \geq 2$, $\sum_n \frac{1}{n^k}$ converges, it shows that there are, in some sense, many more primes than perfect squares. In fact it implies that there is no $\delta < 1$ and constant C such that $\pi(n) \leq Cn^\delta$ for all n, so that if $\pi(n)$ is well-behaved enough to have a "true order of magnitude" than its true order is rather close to n itself.

A striking substance-theoretic result that we will not be able to prove here:

THEOREM 10.8. *(Brun) The set $\mathcal{T}$ of "twin primes" – i.e., primes p for which at least one of $p-2$ and $p+2$ is prime – is insubstantial.*

In a sense, this is disappointing, because we do not know whether $\mathcal{T}$ is infinite, whereas if $\mathcal{T}$ had turned out to be substantial we would immediately know that infinitely many twin primes exist! Nevertheless a fair amount of work has been devoted (for some reason) to calculating **Brun's sum**

$$\sum_{n \in \mathcal{T}} \frac{1}{n} \approx 1.902\ldots.$$

In particular Tom Nicely has done extensive computations of Brun's sum. His work got some unexpected publicity in the mid 1990's when his calculations led to the recognition of the infamous "Pentium bug", a design flaw in many of the Intel microprocessors.[6]

[6]The PC I bought in 1994 (my freshman year of college) had such a bug. The Intel corporation reassured consumers that the bug would be of no practical consequence unless they were doing substantial floating point arithmetic. Wonderful...

The last word on density versus substance: In 1972 Endre Szemerédi proved – by elementary combinatorial means – the sensational result that any subset S of positive upper density contains arbitrarily long arithmetic progressions, a vast generalization of a famous theorem of van der Waerden (on "colorings") which was conjectured by Erdős and Turan in 1936.[7] Unfortunately this great theorem does not apply to the primes, which have zero density.

However, Erdős and Turan made the much more ambitious conjecture that any substantial subset should contain arbitrarily long arithmetic progressions. Thus, when Green and Tao proved in 2002 that there *are* arbitrarily long arithmetic progressions in the primes, they verified a very special case of this conjecture. Doubtless many mathematicians are now reconsidering the Erdős-Turan conjecture with renewed seriousness.

5. Euclid-Mullin Sequences

Let $c \in \mathbb{Z}^+$, $a \in \mathbb{Z}^\bullet$, and let q_1 be a prime number, such that

$$\gcd(a, cq_1) = 1,\ a + cq_1 \geq 2.$$

F$n \in \mathbb{Z}^+$, having chosen primes $q_1, \ldots, q_n$, we take q_{n+1} to be a prime divisor of

$$a + cq_1 \cdots q_n.$$

Note first that $a + cq_1 \cdots q_n \geq a + cq_1 \geq 2$, so such a prime divisor exists. Moreover we assume by induction that $\gcd(a, cq_1 \cdots q_n) = 1$. Then if $q_{n+1} = q_i$ for some $1 \leq i \leq n$ then since $q_i = q_{n+1} \mid a + cq_1 \cdots q_n$, we have $q_i \mid a$: contradiction. So $\gcd(q_1 \cdots q_n, q_{n+1}) = 1$. Thus if $q_{n+1} \mid a$, then $q_{n+1} \mid c$, contradiction. So $\gcd(a, cq_1 \cdots q_{n+1}) = 1$.

For example, take $(q_1, c, a) = (2, 1, 1)$. Then

$$a + cq_1 = 1 + 1 \cdot 2 = 3,\ \text{so } q_2 = 3.$$

$$a + cq_1q_2 = 1 + 1 \cdot 2 \cdot 3 = 7,\ \text{so } q_3 = 7.$$

$$a + cq_1q_2q_3 = 1 + 1 \cdot 2 \cdot 3 \cdot 7 = 43,\ \text{so} q_4 = 43.$$

But:

$$a + cq_1q_2q_3q_4 = 1 + 2 \cdot 3 \cdot 7 \cdot 43 = 13 \cdot 139,$$

and now we see that our above procedure is valid but not completely determinate: we could follow it by taking either $q_5 = 13$ or $q_5 = 139$. In 1963 the *American Mathematical Monthly* published Research Problems, including one by A.A. Mullin. Still in the case $(q_1, c, a) = (2, 1, 1)$, Mullin suggested two recipes for resolving the indeterminacy: in his first sequence, we take q_{n+1} to be the **least** prime divisor of $a + cq_1 \cdots q_n$, and in his second sequence we take q_{n+1} to be the **greatest** prime divisor of $a + cq_1 \cdots q_n$. Later authors have spoken of the **first and second Euclid-Mullin sequences**. These definitions carry over immediately to the case of general (q_1, c, a): taking the least prime divisor we define a sequence of distinct

[7]Several other mathematicians have devoted major parts of their career to bringing more sophisticated technology to bear on this problem, obtaining quantitative improvements of Szemerédi's theorem. Notably Timothy Gowers received the Fields Medal in 1998 for his work in this area. One must wonder whether the fact that Szemerédi did not receive the Fields Medal for his spectacular result is an instance of the prejudice against combinatorial mathematics in the mainstream mathematical community. (The extent of this prejudice also renders the plot of the movie "Good Will Hunting" somewhat implausible.)

primes $\mathrm{EML}_1(q_1, c, a)$, and taking the greatest prime divisor we define a sequence of distinct primes $\mathrm{EML}_2(q_1, c, a)$.

EXAMPLE 10.9. *a) The known terms of* $\mathrm{EML}_1(2, 1, 1)$ *are*[8]

$$2, 3, 7, 43, 13, 53, 5, 6221671, 38709183810571, 139, 2801, 11, 17, 5471, 52662739, 23003,$$
$$30693651606209, 37, 1741, 1313797957, 887, 71, 7127, 109, 23, 97, 159227,$$
$$643679794963466223081509857, 103, 1079990819, 9539, 3143065813, 29, 3847, 89, 19,$$
$$577, 223, 139703, 457, 9649, 61, 4357,$$
$$227432689108589532754984915075774848386671439568260420754414940780761245893,$$
$$59, 31, 211$$

The last four terms were computed in 2012. The sequence behaves very irregularly, and determining when or even whether a given prime occurs in the sequence appears to be very difficult. For instance, only in 2012 did we learn that 31 *appears in* $\mathrm{EML}_1(2, 1, 1,)$*, and at the time of this writing it is not known whether* 41 *appears in the sequence.*

b) The known terms of $\mathrm{EML}_2(2, 1, 1)$ *are ...*

[8]courtesy of `https://oeis.org/A000945` and `http://www.mersenneforum.org/showpost.php?p=311145&postcount=52`

CHAPTER 11

The Prime Number Theorem and the Riemann Hypothesis

1. Some History of the Prime Number Theorem

Recall we have defined, for positive real x,

$$\pi(x) = \#\ \{\text{primes } p \leq x\}.$$

The following is probably the single most important result in number theory.

THEOREM 11.1. *(Prime Number Theorem) We have $\pi(x) \sim \frac{x}{\log x}$; i.e.,*

$$\lim_{x\to\infty} \frac{\pi(x)\log x}{x} = 1.$$

1.1. Gauss at 15. The prime number theorem (affectionately called "PNT") was apparently first conjectured in the late 18th century, by Legendre and Gauss (independently). In particular, Gauss conjectured an equivalent – but more appealing – form of the PNT in 1792, at the age of 15 (!!!).

Namely, he looked at the frequency of primes in intervals of lengths 1000:

$$\Delta(x) = \frac{\pi(x) - \pi(x-1000)}{1000}.$$

Computing by hand, Gauss observed that $\Delta(x)$ seemed to tend to 0, however very slowly. To see how slowly he computed the reciprocal, and found

$$\frac{1}{\Delta(x)} \approx \log x,$$

meaning that

$$\Delta(x) \approx \frac{1}{\log x}.$$

Evidently 15 year old Gauss knew both differential and integral calculus, because he realized that $\Delta(x)$ was a slope of the secant line to the graph of $y = \pi(x)$. When x is large, this suggests that the slope of the tangent line to $\pi(x)$ is close to $\frac{1}{\log x}$, and hence he guessed that the function

$$\mathrm{Li}(x) := \int_2^x \frac{dt}{\log t}$$

was a good approximation to $\pi(x)$.

PROPOSITION 11.2. *We have*

$$\mathrm{Li}(x) \sim \frac{x}{\log x}.$$

PROOF. A calculus exercise (L'Hôpital's rule!). □

Thus PNT is equivalent to $\pi(x) \sim \mathrm{Li}(x)$. The function $\mathrm{Li}(x)$ – called the **logarithmic integral** – is not elementary, but has a simple enough power series expansion (see for yourself). Nowadays we have lots of data, and one can see that the error $|\pi(x) - \mathrm{Li}(x)|$ is in general much smaller than $|\pi(x) - \frac{x}{\log x}|$, so the dilogarithm gives a "better" asymptotic expansion. (How good? Read on.)

1.2. A partial result. As far as I know, there was no real progress for more than fifty years, until the Russian mathematician Pafnuty Chebyshev proved the following two impressive results.

THEOREM 11.3. *(Chebyshev, 1848, 1850)*
a) There exist explicitly computable positive constants C_1, C_2 such that for all x,

$$\frac{C_1 x}{\log x} < \pi(x) < \frac{C_2 x}{\log x}.$$

b) If $\lim_{x\to\infty} \frac{\pi(x)}{x/(\log x)}$ exists, it necessarily equals 1.

Remarks:

(i) For instance, one version of the proof gives $C_1 = 0.92$ and $C_2 = 1.7$.
(But I don't know what values Chebyshev himself derived.)

(ii) The first part shows that $\pi(x)$ is of "order of magnitude" $\frac{x}{\log x}$, and the second shows that if it is "regular enough" to have an asymptotic value at all, then it must be asymptotic to $\frac{x}{\log x}$. Thus the additional trouble in proving PNT is establishing this *regularity* in the distribution of the primes, a quite subtle matter. (We have seen that other arithmetical functions, like φ and d are far less regular than this – their upper and lower orders differ by more than a multiplicative constant, so the fact that this regularity should exist for $\pi(x)$ is by no means assured.)

(iii) Chebyshev's proof is quite elementary: it uses less machinery than some of the other topics in this course. However we will not give the time to prove it here: blame it on your instructor's failure to "understand" the proof.

1.3. A complex approach.

The next step was taken by Riemann in 1859. We have seen the zeta function

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

and its relation to the primes (e.g. obtaining a proof that $\pi(x) \to \infty$ by the above factorization). However, Riemann considered $\zeta(s)$ as a function of a complex variable: $s = \sigma + it$ (indeed he used these rather strange names for the real and imaginary parts in his 1859 paper, and we have kept them ever since), so

$$n^s = n^{\sigma+it} = n^\sigma n^{it}.$$

Here n^σ is a real number and $n^{it} = e^{i(\log n)t}$ is a point on the unit circle, so in modulus we have $|n^s| = n^\sigma$. From this we get that $\zeta(s)$ is *absolutely convergent* for $\sigma = \Re(s) > 1$. Using standard results from analysis, one sees that it indeed defines

an *analytic* function in the half-plane $\sigma > 1$. Riemann got the zeta function named after him by observing the following:

Fact: $\zeta(s)$ extends ("meromorphically") to the entire complex plane and is analytic everywhere except for a simple pole at $s = 1$.

We recall in passing, for those with some familiarity with complex variable theory, that the extension of an analytic function defined in one (connected) domain in the complex plane to a larger (connected) domain is *unique* if it exists at all: this is the *principle of analytic continuation*. So the zeta function is well-defined. The continuation can be shown to exist via an integral representation valid for $\sigma > 0$ and a **functional equation** relating the values of $\zeta(s)$ to that of $\zeta(1 - s)$. (Note that the line $\sigma = \frac{1}{2}$ is fixed under the $s \mapsto 1 - s$.) Riemann conjectured, but could not prove, certain simple (to state!) analytic properties of $\zeta(s)$, which he saw had profound implications on the distribution of the primes.

1.4. A nonvanishing theorem.

It is a testament to the difficulty of the subject that even after this epochal paper the proof of PNT did not come for almost 40 years. In 1896, Jacques Hadamard and Charles de la Vallée-Poussin proved PNT, independently, but by rather similar methods. The key point in both of their proofs (which Riemann could not establish) was that $\zeta(s) \neq 0$ for any $s = 1 + it$, i.e., along the line with $\sigma = 1$.

Their proof does come with an explicit error estimate, albeit an ugly one.

THEOREM 11.4. *There exist positive constants C and a such that*

$$|\pi(x) - \mathrm{Li}(x)| \leq Cxe^{-a\sqrt{\log x}}.$$

It is not completely obvious that this is indeed an error bound, i.e., that

$$\lim_{x\to\infty} \frac{e^{-a\sqrt{\log x}}}{\mathrm{Li}(x)} = 0.$$

This is left as another calculus exercise.

1.5. An elementary proof is prized.

Much was made of the fact that the proof of PNT, a theorem of number theory, used nontrivial results from complex analysis (which by the end of the 19th century had been developed to a large degree of sophistication). Many people speculated on the existence of an "elementary" proof, a yearning that to my knowledge was never formalized precisely. Roughly speaking it means a proof that uses no extraneous concepts from higher analysis (such as complex analytic functions) but only the notion of a limit and the definition of a prime. It thus caused quite a stir when Atle Selberg and Paul Erdős (not independently, but not quite collaboratively either – the story is a controversial one!) gave what all agreed to be an elementary proof of PNT in 1949. In 1950 Selberg (but not Erdős) received the Fields Medal.

In recent times the excitement about the elementary proof has dimmed: most experts agree that it is less illuminating and less natural than the proof via Riemann's zeta function. Moreover the elementary proof remains quite intricate: ironically, more so than the analytic proof for those with some familiarity with functions

of a complex variable. For those who do not, the time taken to learn some complex analysis will probably turn out to be time well spent.

1.6. Equivalents of PNT.

Many statements are "equivalent" to PNT: i.e., it is much easier to show that they imply and are implied by PNT than to prove them. Here's one:

THEOREM 11.5. *Let p_n be the nth prime. Then*

$$p_n \sim n \log n.$$

Note that this result implies (by the integral test) that $\sum_{p \leq n} \frac{1}{p} \sim \log \log n$; strangely this consequence is much easier to prove than PNT itself.

Far more intriguing is that that PNT is equivalent to an asymptotic formula for the average value of the Möbius function:

THEOREM 11.6.

$$\lim_{N \to \infty} \frac{\sum_{n=1}^{N} \mu(n)}{N} = 0.$$

Recall that the Möbius function is 0 if n is not squarefree (which we know occurs with density $1 - \frac{6}{\pi^2}$) and is $(-1)^r$ if n is a product of r distinct primes. We also saw that the set of all positive integers divisible by only a bounded number, say k, of primes is equal to zero, so most integers $1 \leq n \leq N$ are divisible by lots of primes, and by adding up the values of μ we are recording $+1$ if this large number is even and -1 if this large number is odd. It is very tempting to view this parity as being essentially random, similar to what would happen if we flipped a coin for each (squarefree) n and gave ourselves $+1$ if we got heads and -1 if we got tails.

With this "randomness" idea planted in our mind, the above theorem seems to assert that if we flip a large number N of coins then (with large probability) the number of heads minus the number of tails is small compared to the total number of coin flips. But now it seems absolutely crazy that this result is equivalent to PNT since – under the (as yet completely unjustified) assumption of randomness – it is far too weak: doesn't probability theory tell us that the running total of heads minus tails will be likely to be on the order of the **square root** of the number of coin flips? Almost, but not quite. And *is* this probabilistic model justified? Well, that is the $ 1 million dollar question.

2. Coin-Flipping and the Riemann Hypothesis

Let us define the **Mertens function**

$$M(N) = \sum_{n=1}^{N} \mu(n).$$

The goal of this lecture is to discuss the following seemingly innocuous question.

QUESTION 5. *What is the upper order of $M(N)$?*

Among other incentives for studying this question there is a large financial one: if the answer is close to what we think it is, then proving it will earn you $ 1 million!

Recall $\mu(n)$ takes on only the values ± 1 and 0, so the "trivial bound" is

$$M(N) \leq N.$$

In fact we can do better, since we know that $\mu(n) = 0$ iff n is not squarefree, and we know, asymptotically, how often this happens. This leads to an asymptotic expression for the "absolute sum":

$$\sum_{n=1}^{N} |\mu(n)| = \#\{\text{squarefree } n \leq N\} \sim \frac{6}{\pi^2} N.$$

However, in the last lecture we asserted that $\frac{M(N)}{N} \to 0$, which we interpreted as saying that the average order of μ is asymptotically 0. Thus the problem is one of *cancellation* in a series whose terms are sometimes positive and sometimes negative. Stop for a second and recall how much more complicated the theory of "conditional" convergence of such series is than the theory of convergence of series with positive terms. It turns out that the problem of *how much cancellation to expect* in a series whose terms are sometimes positive and sometimes negative (or a complex series in which the arguments of the terms are spread around on the unit circle) is absolutely a fundamental one in analysis and number theory. Indeed in such matters we can draw fundamental inspiration (if not proofs, directly) from **probability theory**, and to do so – i.e., to make heuristic probabilistic reasoning even in apparently "deterministic" situations – is an important theme in modern mathematics ever since the work of Erdős and Kac in the mid 20th century.

But our story starts before the 20th century. In the 1890's Mertens[1] conjectured:

(MC1) $M(N) \leq \sqrt{N}$ for all sufficiently large N.

This is quite bold. As we have seen, in studying orders of magnitude, it is safer to hedge one's bets by at least allowing a multiplicative constant, leading to the weaker

(MC2) $M(N) \leq C\sqrt{N}$ for all N.

The noted Dutch mathematician Stieltjes claimed a proof of (MC2) in 1885. But his proof was never published and was not found among his papers after his death.

It would be interesting to know why Mertens believed (MC1). He did check the inequality for all N up to $N = 10^4$: this is an amusingly small search by contemporary standards. The problem is not as computationally tractable as one might wish, because computing the Möbius function requires factorization of n: that's hard! Nevertheless we now know that (MC1) holds for all $N \leq 10^{14}$.

These are hard problems: while experts have been dubious about (MC1) and (MC2) for over a century, (MC1) was disproved only in 1985.

THEOREM 11.7. *(Odlyzko-te Riele* [**OlR85**]*) There are explicit constants $C_1 > 1$, $C_2 < -1$ such that*

$$\limsup_N \frac{M(N)}{\sqrt{N}} \geq C_1,$$

[1]Franz Mertens, 1840–1927

$$\liminf_N \frac{M(N)}{\sqrt{N}} \leq C_2.$$

In plainer terms, each of the inequalities $-N \leq M(N)$ and $M(N) \leq N$ fails for infinitely many N. The Odlyzko-te Riele proof doesnot supply a concrete value of N for which $M(N) > \sqrt{N}$, but soon after J. Pintz showed [**Pi87**] that $M(N) > \sqrt{N}$ for some $N < e^{3.21 \cdot 10^{64}} \approx 10^{1.4 \cdot 10^{64}}$ (note the double exponential: this is an enormous number!). Recent work of Saouter and te Riele [**SatR14**] shows that this inequality holds for some $N < e^{1.004 \cdot 10^{33}}$. Yet more recently Best and Trudgian have shown that in Theorem 11.7 one may take $C_1 = 1.6383$ $C_2 = -1.6383$ [**BTxx**].

REMARK 11.8. An earlier draft contained the claim that $M(N) > \sqrt{N}$ for some $N < 10^{154}$. I thank Tim Trudgian for bringing this to my attention. Not only is a counterexample to (MC1) for such a "small" value of N not known, in fact it seems quite unlikely that there is a counterexample at anything close to this order of magnitude. Trudgian recommends a work of Kotnik and van de Lune [**KvdL04**] which contains experimental data and conjectures on $M(N)$. In particular they give a conjecture which suggests that the first counterexample to (MC1) should occur at roughly $N \approx 10^{2.3 \cdot 10^{23}}$: i.e., *much* smaller than the known counterexamples and **much** larger than 10^{154}. It should be emphasized that such conjectures are rather speculative, and the literature contains several incompatible such conjectures.

We still do not know whether (MC2) holds – so conceivably Stieltjes was right all along and the victim of some terrible mix up – although I am about to spin a tale to try to persuade you that (MC2) should be almost, but not quite, true.

But first, what about the million dollars?

In the last section we mentioned two interesting "equivalents" of PNT. The following theorem takes things to another level:

THEOREM 11.9. *The following (unknown!) assertions are equivalent:*

a) For all $\epsilon > 0$, there exists a constant C_ϵ such that $|M(N)| \leq C_\epsilon N^{\frac{1}{2}+\epsilon}$.

b) $|\pi(x) - \mathrm{Li}(x)| \leq \frac{1}{8\pi}\sqrt{x}\log x$ for all $x \geq 2657$.

c) Suppose $\zeta(s_0) = 0$ for some s_0 with real part $0 < \Re(s_0) < 1$. Then $\Re(s_0) = \frac{1}{2}$.

We note that the somewhat abstruse part c) – which refers to the behavior of the zeta function in a region which it is not obvious how it is defined – is the **Riemann hypothesis** (**RH**). Thus we care about RH (for instance) because it is equivalent to a wonderful error bound in the Prime Number Theorem.

In 2000 the Clay Math Institute set the Riemann Hypothesis as one of seven $ 1 million prize problems. If you don't know complex analysis, no problem: just prove part a) about the order of magnitude of the partial sums of the Möbius function.

Note that (MC1) (which is false!) $\implies$ (MC2) $\implies$ condition a) of the theorem, so in announcing a proof of (MC2) Stieltjes was announcing a *stronger* result than the Riemann hypothesis, which did not have a million dollar purse in his day but was no less a mathematical holy grail then than now. (So you can decide how

likely it is that Stieltjes's paper got lost in the mail and never found.)

But why should we believe in the Riemann hypothesis anyway? There is some experimental evidence for it – in any rectangle $|t| \leq N$, $0 < \sigma < 1$ the zeta function can have only finitely many zeros (this holds for any function meromorphic on $\mathbb{C}$), so one can "find all the zeros" up to a certain imaginary part, and the fact that all of these zeros lie on the critical line – i.e., have real part $\frac{1}{2}$ – has been experimentally confirmed in a certain range of t. It is also known that there are infinitely many zeros lying on the critical line (Hardy) and that even a positive proportion of them as we go up lie on the critical line (Selberg – as I said, a great mathematician). For various reasons this evidence is rather less than completely convincing.

So let us go back to randomness – suppose μ really were a random variable. What would it do, in all probability?

We can consider instead the random walk on the integers, where we start at 0 and at time i, step to the right with probability $\frac{1}{2}$ and step to the left with probability $\frac{1}{2}$. Formally speaking, our walk is given by an infinite sequence $\{\epsilon_i\}_{i=1}^{\infty}$, each $\epsilon_i = \pm 1$. The set of all such sign sequences, $\{\pm 1\}^{\infty}$ forms in a natural way a probability space (meaning it has a natural measure – but don't worry about the details; just hold on for the ride). Then we define a random variable

$$S(N) = \epsilon_1 + \ldots + \epsilon_n,$$

meaning a function that we can evaluate on any sign sequence, and it tells us where we end up on the integers after N steps. Now the miracle of modern probability theory is that it makes perfect sense to ask what the lim sup of S_N is.

If you've had a course in probability theory (good for you...) you will probably remember that S_N should be no larger than $\sqrt{N}$, more or less. But this seems disappointing, because that is (MC1) (or maybe (MC2)), which feels quite dubious for the partial sums of the Möbius function. But in between Mertens' day and ours probability theory grew up, and we now know that $\sqrt{N}$ is not *exactly* the correct upper bound. Rather, it is given by the following spectacular theorem:

THEOREM 11.10. *(Kolmogorov) With probability* 1*, we have*

$$\limsup_{N \to \infty} \frac{S_N}{\sqrt{2N \log \log N}} = 1.$$

Thus if you flip a fair coin N times, then in all probability there will be infinitely many moments in time when your running tally of heads minus tails is larger than any constant times the square root of the number of flips. (Similarly, and symmetrically, the limit infimum is -1.) So true randomness predicts that (MC2) is false. On the other hand, it predicts that the Riemann Hypothesis is true, since indeed for all $\epsilon > 0$ there exists a constant C_ϵ such that $\sqrt{2 \log \log N} < C_\epsilon N^\epsilon$.

So if we believed in the "true randomness" of μ, we would believe the following

CONJECTURE 11.11. *(Good-Churchhouse* [**GC68**]*)*

$$\limsup_{N \to \infty} \frac{M(N)}{\sqrt{N \log \log N}} < \infty.$$

$$\liminf_{N\to\infty} \frac{M(N)}{\sqrt{N}\log\log N} > -\infty.$$

Later mathematicians have extended these calculations and suggested refinements of Conjecture 11.11. Based in part on numerical experimentation, Kotnik and van de Lune [**KvdL04**] instead conjecture

$$\limsup_{N\to\infty} \frac{M(N)}{\sqrt{N}\log\log\log N} \in (0,\infty).$$

This slower growth informs their estimate of the smallest counterexample to (MC1)).

Just to make sure, this conjecture is still significantly more precise than the $|M(N)| \leq C_\epsilon N^{\frac{1}{2}+\epsilon}$ which is equivalent to the Riemann Hypothesis, making it unclear exactly how much we should pay the person who can prove it: \$ 2 million? Or more??

Kolmogorov's "law of the iterated logarithm," and hence Conjecture 11.11, does not seem to be very well-known outside of probabilistic circles.[2] In searching the literature I found a paper from the 1960's predicting such a "logarithm law" for $M(N)$. More recently I have seen another paper suggesting that perhaps it should be $\sqrt{\log\log\log N}$ instead of $\sqrt{\log\log N}$. To be sure, the Möbius function is clearly *not* random, so one should certainly be provisional in one's beliefs about the precise form of the upper bounds on $M(N)$. The game is really to decide whether the Möbius function is "random enough" to make the Riemann hypothesis true.

Nevertheless the philosophy expressed here is a surprisingly broad and deep one: whenever one meets a sum S_N of N things, each of absolute value 1, and varying in sign (or in argument in the case of complex numbers), one wants to know how much cancellation there is, i.e., how far one can improve upon the trivial bound of $|S_N| \leq N$. The mantra here is that if there is really no extra structure in the summands – i.e., "randomness" – then one should expect $S_N \approx \sqrt{N}$, more or less! More accurately the philosophy has two parts, and the part that expresses that $|S_N|$ should be *no smaller* than $\sqrt{N}$ unless there is hidden structure is an extremely reliable one. An example of hidden structure is $a_n = e^{\frac{2\pi i}{N}}$, when in fact

$$\sum_{n=1}^{n} a_n = 0.$$

But here we have chosen to sum over all of the Nth roots of unity in the complex plane, a special situation. The second part of the philosophy allows us to hope that S_N is not *too much* larger than $\sqrt{N}$. In various contexts, any of $C\sqrt{N}$, $\sqrt{N}\log N$, $N^{\frac{1}{2}+\epsilon}$, or even $N^{1-\delta}$ for some $\delta > 0$, may count as being "not too much larger." So in truth our **philosophy of almost squareroot error** is a little bit vague. But it can be, and has been, a shining light in a dark place,[3] and we will see further instances of such illumination.

[2] I learned about Kolmogorov's theorem from a talk at Harvard given by W. Russell Mann.

[3] When all other lights go out?

CHAPTER 12

The Gauss Circle Problem and the Lattice Point Enumerator

1. Introduction

We wish to study a very classical problem: how many lattice points lie on or inside the circle $x^2 + y^2 = r^2$? Equivalently, for how many pairs $(x, y) \in \mathbb{Z}^2$ do we have $x^2 + y^2 \leq r^2$? Let $L(r)$ denote the number of such pairs.

Upon gathering a bit of data, it becomes apparent that $L(r)$ grows quadratically with r, which leads to consideration of $\frac{L(r)}{r^2}$. Now:

$$L(10)/10^2 = 3.17.$$
$$L(100)/100^2 = 3.1417.$$
$$L(1000)/1000^2 = 3.141549.$$
$$L(10^4)/10^8 = 3.14159053.$$

The pattern is pretty clear!

THEOREM 12.1. *As $r \to \infty$, we have $L(r) \sim \pi r^2$. Explicitly,*

$$\lim_{r\to\infty} \frac{L(r)}{\pi r^2} = 1.$$

Once stated, this result is quite plausible geometrically: suppose that you have to tile an enormous circular bathroom with square tiles of side length 1 cm. The total number of tiles required is going to be very close to the area of the floor in square centimeters. Indeed, starting somewhere in the middle you can do the vast majority of the job without even worrying about the shape of the floor. Only when you come within 1 cm of the boundary do you have to worry about pieces of tiles and so forth. But the number of tiles required to cover the boundary is something like a constant times the perimeter of the region in centimeters – so something like $C\pi r$ – whereas the number of tiles in the interior is close to πr^2. Thus the contribution to the boundary is neglible: precisely, when divided by r^2, b it approaches 0 as $r \to \infty$.

I myself find this heuristic convincing but not quite rigorous. More precisely, I believe it for a circular region and become more concerned as the boundary of the region becomes more irregularly shaped, but the heuristic doesn't single out exactly what nice properties of the circle are being used. Moreover the "error" bound is fuzzy: it would be useful to know an explicit value of C.

To be more quantitative about it, we define the error

$$E(r) = |L(r) - \pi r^2|,$$

so that Theorem 12.1 is equivalent to the statement

$$\lim_{r\to\infty} \frac{E(r)}{r^2} = 0.$$

The above heuristic suggests that $E(r)$ should be bounded above by a linear function of r. The following elementary result was proved by Gauss in 1837.

THEOREM 12.2. *For all $r \geq 7$, $E(r) \leq 10r$.*

PROOF. Let $P = (x, y) \in \mathbb{Z}^2$ be such that $x^2 + y^2 \leq r^2$. To P we associate the square $S(P) = [x, x + 1] \times [y, y + 1]$, i.e., the unit square in the plane which has P as its lower left corner. Note that the diameter of $S(P)$ – i.e., the greatest distance between any two points of $S(P)$ – is $\sqrt{2}$. So, while P lies within the circle of radius r, $S(P)$ may not, but it certainly lies within the circle of radius $r + \sqrt{2}$. It follows that the total area of all the squares $S(P)$ – which is nothing else than the number $L(r)$ of lattice points – is at most the area of the circle of radius $r + \sqrt{2}$, i.e.,

$$L(r) \leq \pi(r + \sqrt{2})^2 = \pi r^2 + 2\sqrt{2}\pi r + 2\pi.$$

A similar argument gives a lower bound for $L(r)$. Namely, if (x, y) is any point with distance from the origin at most $r - \sqrt{2}$, then the entire square $(\lfloor x \rfloor, \lfloor x + 1 \rfloor) \times (\lfloor y \rfloor, \lfloor y + 1 \rfloor)$ lies within the circle of radius r. Thus the union of all the unit squares $S(P)$ attached to lattice points on or inside $x^2 + y^2 = r$ covers the circle of radius $r - \sqrt{2}$, giving

$$L(r) \geq \pi(r - \sqrt{2})^2 = \pi r^2 - 2\sqrt{2}\pi r + 2\pi.$$

Thus

$$E(r) = |L(r) - \pi r^2| \leq 2\pi + 2\sqrt{2}\pi r \leq 7 + 9r \leq 10r,$$

the last inequality holding for all $r \geq 7$. □

This argument skillfully exploits the geometry of the circle. I would like to present an alternate argument with a much different emphasis.

The first step is to notice that instead of counting lattice points in an expanding sequence of closed disks, it is equivalent to fix the plane region once and for all – here, the unit disk $D : x^2+y^2 \leq 1$ – and consider the number of points $(x, y) \in \mathbb{Q}^2$ with $rx, ry \in \mathbb{Z}$. That is, instead of dividing the plane into squares of side length one, we divide it into squares of side length $\frac{1}{r}$. If we now count these "$\frac{1}{r}$-lattice points" inside D, a moment's thought shows that this number is precisely $L(r)$.

What sort of thing is an area? In calculus we learn that areas are associated to integrals. Here we wish to consider the area of the unit disk as a **double integral** over the square $[-1, 1]^2$. In order to do this, we need to integrate the **characteristic function** χ_D of the unit disk: that is, $\chi(P)$ evaluates to 1 if $P \in D$ and $\chi(P) = 0$ otherwise. The division of the square $[-1, 1]^2$ into $4r^2$ subsquares of side length $\frac{1}{r}$ is exactly the sort of sequence of partitions that we need to define a Riemann sum: that is, the maximum diameter of a subrectangle in the partition is $\frac{\sqrt{2}}{r}$, which tends to 0 as $r \to \infty$. Therefore if we choose any point $P^*_{i,j}$ in each subsquare, then

$$\Sigma_r := \frac{1}{r^2} \sum_{i,j} \chi(P^*_{i,j})$$

is a sequence of Riemann sums for χ_D, and thus

$$\lim_{r\to\infty} \Sigma_r = \int_{[-1,1]^2} \chi_D = \text{Area}(D) = \pi.$$

But we observe that Σ_r is very close to the quantity $L(r)$. Namely, if we take each sample point to be the lower left corner of corner of the corresponding square, then $r^2\Sigma_r = L(r) - 2$, because every such sample point is a lattice point (which gets multiplied by 1 iff the point lies inside the unit circle) and the converse is true except that the points $(1,0)$ and $(0,1)$ are not chosen as sample points. So

$$\lim_{r\to\infty} \frac{L(r)}{r^2} = \lim_{r\to\infty} \frac{L(r)-2+2}{r^2} = \lim_{r\to\infty} \Sigma_r + 0 = \pi.$$

The above argument is less elementary than Gauss's and gives a weaker result: no explicit upper bound on $E(r)$ is obtained. So why have we bothered with it? The answer lies in the generality of this latter argument. We can replace the circle by any **plane region** $R \subset [-1,1]^2$. For any $r \in \mathbb{R}^{>0}$, we define the r-dilate of R,

$$rR = \{rP \mid P \in R\}.$$

This is a plane region which is "similar" to R in the usual sense of Euclidean geometry. Note that if $R = D$ is the closed unit disk then $rD = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq r^2\}$ is the closed disk of radius r. Therefore a direct generalization of the counting function $L(r)$ is

$$L_R(r) = \#\{(x,y) \in \mathbb{Z}^2 \cap rR\}.$$

As above, we can essentially view $\frac{L_R(r)}{r^2}$ as a sequence of Riemann sums for $\int_{[-1,1]^2} \chi_R$ – "essentially" because any lattice points with x or y coordinate equal to 1 exactly will contribute to $L_R(r)$ but not to the Riemann sum. But since the total number of $\frac{1}{r}$-squares which touch the top and/or right sides of the square $[-1,1]^2$ is $4r+1$, this discrepancy goes to 0 when divided by r^2. (Another approach is just to assume that R is contained in the interior $(-1,1)^2$ of the unit square. It should be clear that this is no real loss of generality.) We get the following result:

THEOREM 12.3. *Let $R \subset [-1,1]^2$ be a planar region. Then*

$$\lim_{r\to\infty} \frac{L_R(r)}{r^2} = \text{Area}(R). \tag{40}$$

There remains a technical issue: what do we mean by a "plane region"? Any subset of $[-1,1]^2$? A Lebesgue measurable subset? Neither of these is correct: take

$$I = \{(x,y) \in [-1,1]^2 \mid x, y \in \mathbb{R} \setminus \mathbb{Q}\},$$

i.e., the subset of the square $[-1,1]^2$ consisting of points both of whose coordinates are irrational. Then I is obtained by removing from $[-1,1]^2$ a set of Lebesgue measure zero, so I has Lebesgue measure 4 and thus $\int_{[-1,1]^2} \chi_I$ exists in the Lebesgue sense and is equal to 4. On the other hand, I contains no rational points whatsoever, so for all $r \in \mathbb{Z}^+$, $L_R(r) = 0$. Thus, if we restrict r to positive integral values, then both sides of (40) are well-defined, but they are unequal: $0 \neq 4$.

Looking back at the argument, what is needed is precisely the **Riemann integrability** of the characteristic function χ_D of the region D. It is a basic fact that a bounded function on a bounded domain is Riemann integrable if and only if it is continuous except on a set of measure zero. The characteristic function χ_D is

discontinuous precisely along the boundary of D, so the necessary condition on D is that its boundary have measure zero. (Explicitly, this means that for any $\epsilon > 0$, there exists an infinite sequence R_i of open rectangles whose union covers D and such that the sum of the areas of the rectangles is less than or equal to ϵ.) In geometric measure theory, such regions are called **Jordan measurable**, and this is the condition we need on our "planar region".

Jordan measurability is a relatively mild condition: for instance any region bounded by a piecewise smooth curve (a circle, ellipse, polygon...) is Jordan measurable. In fact a large collection of regions with fractal boundaries are Jordan measurable: for instance Theorem 12.3 applies with R a copy of the **Koch snowflake**, whose boundary is a nowhere differentiable curve.

2. Better Bounds

2.1. The soft/hard dichotomy. As in the previous section, suppose we have a plane region $R \subset [-1,1]^2$, and consider the function $L_R(r)$ which counts the number of lattice points in the dilate rR of R. The main qualitative, or **soft**, result of the last section was

$$L_R(r) \sim \text{Area}(R) r^2.$$

But if we take a more *quantitative*, or **hard**, view, this is only the beginning. Namely, as before, we define

$$E_R(r) = |L_R(r) - \text{Area}(R) r^2|.$$

Theorem 12.3 tells us that $\lim_{r \to \infty} E_R(r) = 0$: this is a fundamentally soft-analytic result. A hard-analytic result would give an explicit upper bound on $E_R(r)$. Theorem 12.1 does just this, albeit in the special case where R is the closed unit disk:

$$E_R(r) \leq 10r.$$

Here are some natural questions:

QUESTION 6. *(Gauss's Circle Problem) In the case of $R = D$, how much can one improve on Gauss's bound $E_D(r) \leq 10r$? Can we find nontrivial lower bounds? What is the "truth" about $E_D(r)$?*

QUESTION 7. *Can one give an explicit upper bound on $E_R(r)$ for an arbitrary plane region R? Could we have, for instance, that $E_r(R)$ is always bounded by a linear function of r? Or by an even smaller power of r?*

Question 6 has received much attention over the years. Let's look again at the data:

$$r = 10: \ L(r) = 317, \ \pi r^2 \approx 314, \ E(r) = 2.8407\ldots$$

$$r = 100: \ L(r) = 31417, \ \pi r^2 \approx 31415.926, \ E(r) = 1.0734\ldots$$

$$r = 1000: \ L(r) = 3141549, \pi r^2 \approx 3141592.653, \ E(r) = 43.653\ldots$$

$$r = 10000: \ L(r) = 314159053, \ \pi r^2 \approx 314159265.358, \ E(r) = 212.3589\ldots$$

We now attempt to describe $E(r)$ by a **power law**: i.e., to find a real number α such that $E(r)$ grows like r^α. If $E(r) \approx r^\alpha$, then $\log E(r) \approx \alpha \log r$, so that to test for a power law we should consider the ratio $P(r) := \frac{\log E(r)}{\log r}$ and see whether it

tends towards some constant α as $r \to \infty$. We have at the moment only four values of r, so this is quite rough, but nevertheless let's try it:

$$r = 10 : P(r) = .453\ldots,$$
$$r = 100 : P(r) = .01538\ldots,$$
$$r = 1000 : P(r) = .54667\ldots,$$
$$r = 10000 : P(r) = .5817\ldots.$$

Whatever is happening is happening quite slowly, but it certainly seems like $E(r) \leq Cr^{\alpha}$ for some α which is safely less than 1.

The first theoretical progress was made in 1904 by a Polish undergraduate, in competition for a prize essay sponsored by the departments of mathematics and physics at the University of Warsaw. The student showed that there exists a constant C such that $E_D(r) \leq Cr^{\frac{2}{3}}$. His name was **Waclaw Sierpinski**, and this was the beginning of a glorious mathematical career.[1]

On the other hand, in 1916 G.H. Hardy and E. Landau, independently, proved that *there does not exist* a constant C such that $E(r) \leq Cr^{\frac{1}{2}}$. The conventional wisdom however says that $r^{\frac{1}{2}}$ is very close to the truth: namely, it is believed that for every real number $\epsilon > 0$, there exists a constant C_ϵ such that

$$E(r) \leq C_\epsilon r^{\frac{1}{2}+\epsilon}. \tag{41}$$

Remark: It is not hard to show that this conjecture implies that

$$\lim_{r\to\infty} P(r) = \lim_{r\to\infty} \frac{\log E(r)}{\log r} = \frac{1}{2}.$$

(Note that the calculations above certainly are not sufficient to suggest this result. It would therefore be interesting to extend these calculations and see if convergence to $\frac{1}{2}$ becomes more apparent.)

Note that Theorem 12.1 above tells us we can take $\epsilon = \frac{1}{2}$ and $C_\epsilon = 10$, whereas Sierpinski showed that we can take $\epsilon = \frac{1}{6}$.[2] The best current bound was proven by Huxley in 1993: he showed that (41) holds for every $\epsilon > 19/146 = 0.13\ldots$. In early 2007 a preprint of Cappell and Shaneson appeared on the arxiv, which claims to establish (41) for every $\epsilon > 0$. As of this writing (Spring of 2009) the paper has not been published, nor do I know any expert opinion on its correctness.

As for Question 7, we begin with the following simple but enlightening example.

Example: Let $R = [-1, 1]^2$ be the square of sidelength 2 centered at the origin. Then $\text{Area}(R) = 4$, so that for any $r \in \mathbb{R}^+$ we have $\text{Area}(rR) = 4r^2$. On the other hand, for $r \in \mathbb{Z}^+$, we can determine $L_R(r)$, the number of lattice points in $[-r, r]^2$

[1] Sierpinski (1882-1969) may well be the greatest Polish mathematician of all time, and Poland is a country with an especially distinguished mathematical tradition. Sierpinski is most remembered nowadays for the fractal triangle pattern that bears his name. I have encountered his work several times over the years, and the work on the Gauss Circle Problem is typical of his style: his theorems have elementary but striking statements and difficult, intricate proofs.

[2] I don't know what value he had for $C_{\frac{1}{3}}$ or even whether his proof gave an explicit value.

exactly: there are $2r+1$ possible values for the x coordinate and the same number of possible values for the y-coordinate, so that $L_r(R) = (2r+1)^2 = 4r^2 + 4r + 1$. In this case we have

$$E_R(r) = |L_r(R) - \text{Area}(rR)| = 4r + 1,$$

so that the true error is a linear function of r. This makes us appreciate Sierpinski's result more: to get a bound of $E_D(r) \leq Cr^\alpha$ for some $\alpha < 1$ one does need to use properties specific to the circle: in the roughest possible terms, there cannot be as many lattice points on the boundary of a curved region as on a straight line segment.

More formally, in his 1919 thesis van der Corput proved the following result:

THEOREM 12.4. *Let $R \subset \mathbb{R}^2$ be a bounded planar region whose boundary is C^∞-smooth and with nowhere vanishing curvature. Then there exists a constant C (depending on R) such that for all sufficiently large $r \in \mathbb{R}^{>0}$,*

$$|L_R(r) - \text{Area}(R)r^2| \leq Cr^{\frac{2}{3}}.$$

It is also known that this result is best possible – there are examples of regions with very nice boundaries in which the power $\frac{2}{3}$ cannot be lowered. (Thus again, the circle is very special!) There are many results which study how the behavior of the error term depends on the assumptions one makes about the boundary of R. To go to the other extreme, a 1997 result of L. Colzani shows that for any bounded region R whose boundary has fractal dimension at most α (this has a technical meaning particular to Colzani's paper; we do not give an explicit definition here), then

$$|L_r(R) - \text{Area}(R)r^2| \leq Cr^{2-\alpha}.$$

As far as I know, it is an **open problem** to give corresponding lower bounds: for instance, to construct, for any $\epsilon > 0$, a region R such that $|L_r(R) - \text{Area}(R)r^2| > r^{2-\epsilon}$ for infinitely many positive integers r. (I myself do not know how to construct such a region for *any* $\epsilon < 1$.)

3. Connections to average values

The reader may well be wondering why the Gauss Circle Problem counts as number theory. On the one hand, as we will see later on, number theory is very much concerned with counting lattice points in certain planar and spatial reasons. But more specifically, Gauss' Circle Problem has to do with the average value of an arithmetical function.

Namely, define $r_2(n)$ to be the function which counts the number of $(x, y) \in \mathbb{Z}^2$ such that $n = x^2 + y^2$. The Full Two Squares Theorem says that $r_2(n) > 0$ iff $2 \mid \text{ord}_p(n)$ for every $p \equiv 3 \pmod 4$. As you have seen in the homework, in practice this condition behaves quite erratically. Certainly the function $r_2(n)$ does not have any nice limiting behavior at $n \to \infty$: on the one hand it is 0 infinitely often, and on the other hand it assumes arbitrarily large values.

Much more regularity is displayed by the function $r_2(n)$ "on average." Namely, for any function $f : \mathbb{Z}^+ \to \mathbb{C}$, we define a new function

$$f_{\text{ave}} : n \mapsto \frac{1}{n}\left(f(1) + \ldots + f(n)\right).$$

As its name suggests, $f_{\text{ave}}(n)$ is the average of the first n values of f.

It is also convenient to work also with the summatory function $F(n) := \sum_{k=1}^{n} f(k)$. The relation between them is simple:

$$F(n) = n \cdot f_{\text{ave}}(n).$$

THEOREM 12.5. *For $f(n) = r_2(n)$, $F(n) \sim \pi n$ and $f_{\text{ave}}(n) \sim \pi$.*

PROOF. Indeed, $F(n) = r_2(1) + \ldots + r_2(n)$ counts the number of lattice points on or inside the circle $x^2 + y^2 \leq n$, excepting the origin. Therefore

$$F(n) = L_D(\sqrt{n}) - 1 \sim \pi(\sqrt{n})^2 - 1 \sim \pi n.$$

□

In this context, the Gauss Circle Problem is equivalent to studying the error between $F(n)$ and πn. Studying errors in asymptotic expansions for arithmetic functions is one of the core topics of analytic number theory.

We remark with amusement that the average value of $r_2(n)$ is asymptotically constant and equal to the irrational number π: there is no n for which $r_2(n) = \pi$!

In fact there is a phenomenon here that we should take seriously. A natural question is how often is $r_2(n) = 0$? We know that $r_2(n) = 0$ for all $n = 4k + 3$, so it is equal to zero at least $\frac{1}{4}$ of the time. But the average value computation allows us to do better. Suppose that there exists a number $0 < \alpha \leq 1$ such that $r_2(n) = 0$ at most α proportion of the time. Then $r_2(n) > 0$ at least $1 - \alpha$ of the time, so the average value of $r_2(n)$ is at least $8(1 - \alpha)$. Then $\pi \geq 8(1 - \alpha)$, or

$$\alpha \geq 1 - \pi/8 \approx .607.$$

That is, we've shown that $r_2(n) = 0$ more than 60% of the time.[3]

In fact this only hints at the truth. In reality, $r_2(n)$ is equal to zero "with probability one". In other words, if we pick a large number N and choose at random an elment $1 \leq n \leq N$, then the probability that n is a sum of two squares approaches 0 as $N \to \infty$. This exposes one of the weaknesses of the arithmetic mean (one that those who compose and grade exams become well aware of): without further assumptions it is unwarranted to assume that the mean value is a "typical" value in any reasonable sense. To better capture the notion of typicality one can import further statistical methods and study the **normal order** of an arithmetic function. With regret, we shall have to pass this concept over entirely as being too delicate for our course. See for instance G. Tenenbaum's text [**T**] for an excellent treatment.

The lattice point counting argument generalizes straightforwardly (but fruitfully) to higher-dimensional Euclidean space $\mathbb{R}^N$. For instance, the analogous argument involving lattice points on or inside the sphere of radius r in $\mathbb{R}^3$ gives:

THEOREM 12.6. *The number $R_3(r)$ of integer solutions (x, y, z) to $x^2+y^2+z^2 \leq r^2$ is asymptotic to $\frac{4}{3}\pi r^3$, with error being bounded by a constant times r^2.*

[3]This argument was not intended to be completely rigorous, and it isn't. What it really shows is that it is *not* the case that $r_2(n) = 0$ on a set of density at least $\alpha = 1 - \pi/8$. But this is morally the right conclusion: see below.

COROLLARY 12.7. *The average value of the function $r_3(n)$, which counts representations of n by sums of three integer squares, is asymptotic to $\frac{4}{3}\pi\sqrt{n}$.*

We can similarly give asymptotic expressions for the average value of $r_k(n)$ – the number of representations of n as a sum of k squares – for any k, given a formula for the volume of the unit ball in $\mathbb{R}^k$. We leave it to the reader to find such a formula and thereby compute an asymptotic for the average value of e.g. $r_4(n)$.

CHAPTER 13

Minkowski's Convex Body Theorem

1. The Convex Body Theorem

1.1. Introduction.

We have already considered instances of the following type of problem: given a bounded subset Ω of Euclidean space $\mathbb{R}^N$, to determine $\#(\Omega \cap \mathbb{Z}^N)$, the number of integral points in Ω. It is clear however that there is no answer to the problem in this level of generality: an arbitrary Ω can have any number of lattice points whatsoever, including none at all.

In the pervious chapter we counted lattice points not just on Ω itself but on dilates $r\Omega$ of Ω by positive integers r. We found that for any "reasonable" Ω,

$$L_\Omega(r) := \#(r\Omega \cap \mathbb{Z}^N) \sim r^N \operatorname{Vol}(\Omega). \tag{42}$$

More precisely, we showed that this holds for all bounded sets Ω which are **Jordan measurable**, meaning that the characteristic function $\mathbf{1}_\Omega$ is Riemann integrable.

It is also natural to ask for sufficient conditions on a bounded subset Ω for it to have lattice points at all. One of the first results of this kind is a theorem of Minkowski, which is both beautiful in its own right and indispensably useful in the development of modern number theory (in several different ways).

Before stating the theorem, we need a bit of terminology. Recall that a subset $\Omega \subset \mathbb{R}^N$ is **convex** if for all pairs of points $P, Q \in \Omega$, also the entire line segment

$$\overline{PQ} = \{(1-t)P + tQ \mid 0 \le t \le 1\}$$

is contained in Ω. A subset $\Omega \subset \mathbb{R}^N$ is **centrally symmetric** if whenever it contains a point $v \in \mathbb{R}^N$ it also contains $-v$, the reflection of v through the origin.

A **convex body** is a nonempty, bounded, centrally symmetric convex set.

Some simple observations and examples:

i) A subset of $\mathbb{R}$ is convex iff it is an interval.
ii) A regular polygon together with its interior is a convex subset of $\mathbb{R}^2$.
iii) An open or closed disk is a convex subset of $\mathbb{R}^2$.
iv) Similarly, an open or closed ball is a convex subset of $\mathbb{R}^N$.
v) If Ω is a convex body, then $\exists\ P \in \Omega$; then $-P \in \Omega$ and $0 = \frac{1}{2}P + \frac{1}{2}(-P) \in \Omega$.
vi) The open and closed balls of radius r with center P are convex bodies iff $P = 0$.

Warning: The term "convex body" often has a similar but slightly different meaning: e.g., according to Wikipedia, a convex body is a closed, bounded convex subset Ω of $\mathbb{R}^N$ which has nonempty interior (i.e., there exists at least one point P of Ω such that for sufficiently small $\epsilon > 0$ the entire open ball $B_\epsilon(P)$ of points of $\mathbb{R}^N$ of distance less than ϵ from P is contained in Ω). Our definition of convex body is chosen so as to make the statement of Minkowski's Theorem as clean as possible.

First we record a purely technical result, without proof:

LEMMA 13.1. *(Minkowski) A bounded convex set $\Omega \subset \mathbb{R}^N$ is Jordan measurable: that is, the function*

$$\mathbf{1}_\Omega : x \mapsto 1, x \in \Omega;\ 0, x \notin \Omega$$

*is Riemann integrable. Therefore we can define the **volume** of Ω as*

$$\mathrm{Vol}(\Omega) = \int_{\mathbb{R}^N} \mathbf{1}_\Omega.$$

Here we are using "volume" as a generic term independent of dimensions. When $N = 1$ it would be more properly called "length"; when $N = 2$, "area"; and, perhaps, "hyper-volume" when $N > 3$.

Intuitively speaking, this just says that the boundary of a convex set is not pathologically rugged. In our applications, our bodies will be things like polyhedra and spheres, which are evidently not pathological in this way.

We will also need the following simple result, which ought to be familiar from a course in geometry, multi-variable calculus and/or linear algebra. The reader might try to prove it for herself, but we will not assign it as a formal exercise because we will discuss a more general result in §1.4.

LEMMA 13.2. *(Dilation Lemma) Recall that for a subset Ω of $\mathbb{R}^N$ and a positive real number α we define the **dilate** of Ω*

$$\alpha\Omega := \{\alpha \cdot P = (\alpha x_1, \ldots, \alpha x_n) \mid P = (x_1, \ldots, x_n) \in \Omega\}.$$

Then:
a) Ω is nonempty $\iff$ $\alpha\Omega$ is nonempty.
b) Ω is bounded $\iff$ $\alpha\Omega$ is bounded.
c) Ω is Jordan measurable $\iff$ $\alpha\Omega$ is Jordan measurable, and if so,

$$\mathrm{Vol}(\alpha\Omega) = \alpha^N \mathrm{Vol}(\Omega).$$

d) Ω is convex $\iff$ $\alpha\Omega$ is convex.
e) Ω is centrally symmetric $\iff$ $\alpha\Omega$ is centrally symmetric.

An immediate consequence is:

COROLLARY 13.3. *If $\Omega \subset \mathbb{R}^N$ is a convex body of volume V, then for any positive real number α, $\alpha\Omega$ is a convex body of volume $\alpha^N V$.*

We saw above that any convex body $\Omega \subset \mathbb{R}^N$ contains the origin. In particular, such a set contains at least one point in $\mathbb{Z}^N$. Must it contain any more?

Of course not. Take in the plane the disk of radius r centered at the origin. This is a convex body which, if $r < 1$, does not intersect any other lattice point besides

0. If $r = 1$, it meets the four closest points to 0 if the disk is closed but not if it is open; for $r > 1$ it necessarily meets other lattice points.

Can we find a convex body in $\mathbb{R}^2$ which contains no nonzero lattice points but has larger area than the open unit disk, i.e., area larger than π? Of course we can: the open square

$$(-1,1)^2 = \{(x,y) \in \mathbb{R}^2 \mid |x|, |y| < 1\}$$

has area 4 but meets no nonzero lattice points. As in the case of circles, this is certainly the limiting case *of its kind*: any centrally symmetric – i.e., with vertices $(\pm a, \pm b)$ for positive real numbers a, b – will contain the lattice point $(1,0)$ if $a > 1$ and the lattice point $(0,1)$ if $b > 1$, so if it does not contain any nonzero lattice points we have $\max(a,b) \leq 1$ and thus its area is at most 4. But what if we rotated the rectangle? Or took a more elaborate convex body?

A symmetric convex subset of the real line $\mathbb{R}^1$ is just an interval, either of the form $(-a,a)$ or $[-a,a]$. Thus by reasoning similar to, but even easier than, the above we see that a centrally symmetric convex subset of $\mathbb{R}$ must have a nontrivial lattice point if its "one dimensional volume" is greater than 2, and a centrally symmetric convex *body* (i.e., closed) must have a nontrivial lattice point if its one-dimensional volume is at least 2.

Now passing to higher dimensions, we see that the open cube $(-1,1)^N$ is a symmetric convex subset of volume 2^N which meets no nontrivial lattice point, whereas for any $0 < V < 2^N$ the convex body $[-\frac{V^{1/N}}{2}, \frac{V^{1/N}}{2}]^N$ meets no nontrivial lattice point and has volume V. After some further experimentation, it is natural to suspect the following result.

THEOREM 13.4. *(Minkowski's Convex Body Theorem) Suppose $\Omega \subset \mathbb{R}^N$ is a convex body with $\mathrm{Vol}(\Omega) > 2^N$. Then there exist integers $x_1, \ldots, x_N$, not all zero, such that $P = (x_1, \ldots, x_N) \in \Omega$.*

1.2. First Proof of Minkowski's Convex Body Theorem.

Step 0: By Corollary 13.3, $\frac{1}{2}\Omega$ is also a convex body of volume

$$\mathrm{Vol}(\frac{1}{2}\Omega) = \frac{1}{2^N}\mathrm{Vol}(\Omega) > 1.$$

Moreover Ω contains a nonzero "integral point" $P \in \mathbb{Z}^N$ iff $\frac{1}{2}\Omega$ contains a nonzero "half-integral point" – a nonzero P such that $2P \in \mathbb{Z}^N$. So it suffices to show: for any convex body $\Omega \subset \mathbb{R}^N$ with volume greater than one, there exist integers $x_1, \ldots, x_N$, not all zero, such that $P = (\frac{x_1}{2}, \ldots, \frac{x_N}{2})$ lies in Ω.

Step 1: Observe that if Ω contains P and Q, by central symmetry it contains $-Q$ and then by convexity it contains $\frac{1}{2}P + \frac{1}{2}(-Q) = \frac{1}{2}P - \frac{1}{2}Q$.

Step 2: For a positive integer r, let $L(r)$ be the number of $\frac{1}{r}$-lattice points of Ω, i.e., points $P \in \mathbb{R}^N \cap \Omega$ such that $rP \in \mathbb{Z}^N$. By Lemma 13.1, Ω is Jordan measurable, and then by Theorem 12.3 we have $\lim_{r\to\infty} \frac{L(r)}{r^N} = \mathrm{Vol}(\Omega)$. Since $\mathrm{Vol}(\Omega) > 1$, for sufficiently large r we must have $L(r) > r^N$. Because $\#(\mathbb{Z}/r\mathbb{Z})^N = r^N$, by the

pigeonhole principle there exist distinct integral points

$$P = (x_1, \ldots, x_N) \neq Q = (y_1, \ldots, y_N)$$

such that $\frac{1}{r}P, \frac{1}{r}Q \in \Omega$ and $x_i \equiv y_i \pmod{r}$ for all i. By Step 1 Ω contains

$$R := \frac{1}{2}\left(\frac{1}{r}P\right) - \frac{1}{2}\left(\frac{1}{r}Q\right) = \frac{1}{2}\left(\frac{x_1 - y_1}{r}, \ldots, \frac{x_N - y_N}{r}\right).$$

But $x_i \equiv y_i \pmod{r}$ for all i and therefore $\frac{1}{r}(P - Q) = (\frac{x_1 - y_1}{r}, \ldots, \frac{x_N - y_N}{r}) \in \mathbb{Z}^N$ and thus $R = \frac{1}{2}(\frac{1}{r}(P - Q))$ is a half integral point lying in Ω: QED!

1.3. Second Proof of Minkowski's Convex Body Theorem.

We first introduce some further terminology.

Let $\Omega \subset \mathbb{R}^N$ be a bounded Jordan measurable set. Consider the following set

$$P(\Omega) := \bigcup_{x \in \mathbb{Z}^N} x + \Omega;$$

that is, $P(\Omega)$ is the union of the translates of Ω by all integer points x. We say that Ω is **packable** if the translates are pairwise disjoint, i.e., if for all $x \neq y \in \mathbb{Z}^N$, $(x + \Omega) \cap (y + \Omega) = \emptyset$.

EXAMPLE 13.5. *Let $\Omega = B_0(r)$ be the open disk in $\mathbb{R}^N$ centered at the origin with radius r. Then Ω is packable iff $r \leq \frac{1}{2}$.*

EXAMPLE 13.6. *For $r > 0$, let $\Omega = [0, r]^N$ be the cube with side length r and one vertex on the origin. Then Ω if packable iff $r < 1$, i.e., iff $\mathrm{Vol}(\Omega) < 1$. Also the open cube $(0, 1)^N$ is packable and of volume one.*

These examples serve to motivate the following result.

THEOREM 13.7. *(Blichfeldt's Theorem) If a bounded, Jordan measurable subset $\Omega \subset \mathbb{R}^N$ is packable, then $\mathrm{Vol}(\Omega) \leq 1$.*

PROOF. Suppose that Ω is packable, i.e., that the translates $\{x + \Omega \mid x \in \mathbb{Z}^N\}$ are pairwise disjoint. Let $d > 0$ be such that every point of Ω lies at a distance at most d from the origin.

Let $\overline{B}_r(0)$ be the closed ball of radius r centered at the origin. It has volume $c(N)r^N$ where $c(N)$ depends only on N.[1] By our work on Gauss's Circle Problem, we know that the number of lattice points inside $\overline{B}_r(0)$ is asymptotic to $c(N)r^N$. Therefore the number of lattice points inside $\overline{B}_{r-d}(0)$ is asymptotic, as $r \to \infty$, to $c(N)(r - d)^N \sim c(N)r^N$. Therefore for any fixed $\epsilon > 0$, there exists R such that $r \geq R$ implies that the number of lattice points inside $\overline{B}_{r-d}(0)$ is at least $(1 - \epsilon)c(N)r^N$.

Now note that if $x \in \mathbb{Z}^N$ is such that $||x|| \leq r - d$, then the triangle inequality gives $x + \Omega \subset \overline{B}_0(r)$. Then, if Ω is packable, then we have at least $(1 - \epsilon)c(N)r^N$ pairwise disjoint translates of Ω contained inside $\overline{B}_0(r)$. Therefore we have

$$c(N)r^N = \mathrm{Vol}(\overline{B}_r(0)) \geq \mathrm{Vol}(P(\Omega) \cap \overline{B}_r(0)) \geq (1 - \epsilon)c(N)r^N \,\mathrm{Vol}(\Omega),$$

[1]The values of $c(N)$ are known – of course $c(2) = \pi$ and $c(3) = \frac{4\pi}{3}$ are familiar from our mathematical childhood, and later on you will be asked to compute $c(4) = \frac{\pi^2}{2}$. But as you will shortly see, it would be pointless to substitute in the exact value of $c(N)$ here.

and therefore

$$\mathrm{Vol}(\Omega) \leq \frac{1}{1-\epsilon}.$$

Since this holds for all $\epsilon > 0$, we conclude $\mathrm{Vol}(\Omega) \leq 1$. □

The reader who knows about such things will see that the proof works verbatim if Ω is merely assumed to be bounded and Lebesgue measurable.

Now we use Blichfeldt's Theorem to give a shorter proof of Minkowski's Theorem. As in the first proof, after the rescaling $\Omega \mapsto \frac{1}{2}\Omega$, our hypothesis is that Ω is a convex body with $\mathrm{Vol}(\Omega) > 1$ and we want to prove that Ω contains a nonzero point with half-integral coordinates. Applying Blichfeldt's Lemma to Ω, we get $x, y \in \mathbb{Z}^N$ such that $(x + \Omega) \cap (y + \Omega)$ is nonempty. In other words, there exist $P, Q \in \Omega$ such that $x + P = y + Q$, or $P - Q = y - x \in \mathbb{Z}^N$. But as we saw above, any convex body which contains two points P and Q also contains $-Q$ and therefore $\frac{1}{2}P - \frac{1}{2}Q = \frac{1}{2}(P - Q)$, which is a half-integral point.

1.4. Minkowski's Theorem Mark II.

Let $\Omega \subset \mathbb{R}^N$. In the last section we considered the effect of a dilation on Ω: we got another subset $\alpha\Omega$, which was convex iff Ω was, centrally symmetric iff Ω was, and whose area was related to Ω in a predictable way.

Note that dilation by $\alpha \in \mathbb{R}^{>0}$ can be viewed as a **linear automorphism** of $\mathbb{R}^N$: that is, the map $(x_1, \ldots, x_n) \mapsto (\alpha x_1, \ldots, \alpha x_n)$ is an invertible linear map. Its action on the standard basis $e_1, \ldots, e_N$ of $\mathbb{R}^N$ is simply $e_i \mapsto \alpha e_i$, so its matrix representation is

$$\alpha : \mathbb{R}^N \to \mathbb{R}^N, (x_1, \ldots, x_n)^t \mapsto \begin{bmatrix} \alpha & 0 & 0 & \ldots & 0 \\ 0 & \alpha & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & \alpha \end{bmatrix} (x_1, \ldots, x_n)^t.$$

Now consider a more general linear automorphism $M : \mathbb{R}^N \to \mathbb{R}^N$, which we may identify with its defining matrix $M \in \mathrm{GL}_N(\mathbb{R})$ (i.e., $M = (m_{ij})$ is an $N \times N$ real matrix with nonzero determinant). We will now state – and prove – the following generalization of the dilation lemma to arbitrary linear automorphisms:

Lemma 13.8. *Let Ω be a subset of $\mathbb{R}^N$ and $M : \mathbb{R}^N \to \mathbb{R}^N$ be an invertible linear map. Consider the image*

$$M(\Omega) = \{M(x_1, \ldots, x_n)^t \mid (x_1, \ldots, x_n) \in \Omega\}.$$

a) Ω is nonempty $\iff$ $M(\Omega)$ is nonempty.
b) Ω is bounded $\iff$ $M(\Omega)$ is bounded.
c) Ω is convex $\iff$ $M(\Omega)$ is convex.
d) Ω is centrally symmetric $\iff$ $M(\Omega)$ is centrally symmetric.
e) Ω is Jordan measurable $\iff$ $M(\Omega)$ is Jordan measurable, and if so,

$$\mathrm{Vol}(M(\Omega)) = |\det(M)| \mathrm{Vol}(\Omega).$$

Proof. Part a) is quite obvious. Part b) holds with M replaced by any homeomorphism of $\mathbb{R}^N$: i.e., a continuous map from $\mathbb{R}^N$ to itself with continuous inverse,

because a subset of $\mathbb{R}^N$ is bounded iff it is contained in a compact subset, and the image of a compact subset under a continuous function is bounded. Part c) is true because the image of a line segment under a linear map is a line segment. Part d) follows because of the property $M(-v) = -Mv$ of linear maps. As for part e), the preservation of Jordan measurability follows from the fact that an image of a set of measure zero under a linear map has measure zero. The statement about areas is precisely what one gets by applying the change of variables $(x_1, \ldots, x_N) \mapsto (y_1, \ldots, y_N) = M(x_1, \ldots, x_N)$ in the integral $\int_{\mathbb{R}^N} \mathbf{1} dx_1 \cdots dx_N$. □

COROLLARY 13.9. *If $\Omega \subset \mathbb{R}^N$ is a convex body and $M : R^N \to R^N$ is an invertible linear map, then $M(\Omega)$ is a convex body, and* $\mathrm{Vol}(M(\Omega)) = |\det(M)| \mathrm{Vol}(\Omega)$.

Recall that the lattice points inside $r\Omega$ are precisely the $\frac{1}{r}$-lattice points inside Ω. This generalizes to arbitrary transformations as follows: for $M \in \mathrm{GL}_N(\mathbb{R})$, put

$$\Lambda := M\mathbb{Z}^N = \{M(x_1, \ldots, x_N)^t \mid (x_1, \ldots, x_N) \in \mathbb{Z}^N.$$

The map $\Lambda : \mathbb{Z}^N \to A\mathbb{Z}^N$ is an isomorphism of groups, so $A\mathbb{Z}^N$ is, abstractly, simply another copy of $\mathbb{Z}^N$. However, it is embedded inside $\mathbb{R}^N$ differently. A nice geometric way to look at it is that $\mathbb{Z}^N$ is the vertex set of a tiling of $\mathbb{R}^N$ by unit (hyper)cubes, whereas Λ is the vertex set of a tiling of $\mathbb{R}^N$ by (hyper)parallelopipeds. A single parallelopiped is called a **fundamental domain** for Λ, and the volume of a fundamental domain is given by $|\det(M)|$.[2] We sometimes refer to the volume of the fundamental domain as simply the volume of Λ and write

$$\mathrm{Vol}(\Lambda) = |\det(M)|.$$

Now the fundamental fact – a sort of "figure-ground" observation – is the following:

PROPOSITION 13.10. *Let $\Omega \subset \mathbb{R}^N$ and let $M : \mathbb{R}^N \to \mathbb{R}^N$ be an invertible linear map. Then M induces a bijection between $M^{-1}(\mathbb{Z}^N) \cap \Omega$ and $\mathbb{Z}^N \cap M(\Omega)$.*

If the statement is understood, the proof is immediate!

Applying this (with M^{-1} in place of M) gives the following: if we have a lattice $\Lambda = M\mathbb{Z}^N$, and a convex body Ω, the number of points of $\Lambda \cap \Omega$ is the same as the number of points of $\mathbb{Z}^N \cap M^{-1}(\Omega)$. Since

$$\mathrm{Vol}(M^{-1}(\Omega)) = |\det(M^{-1})| \mathrm{Vol}(\Omega) = \frac{\mathrm{Vol}(\Omega)}{\det(M)} = \frac{\mathrm{Vol}(\Omega)}{\mathrm{Vol}(\Lambda_M)},$$

we immediately deduce a more general version of Minkowski's Theorem.

THEOREM 13.11. *(Minkowski's Theorem Mark II) Let $\Omega \subset \mathbb{R}^N$ be a convex body. Let $M : \mathbb{R}^N \to \mathbb{R}^N$ be an invertible linear map, and put $\Lambda_M = M(\mathbb{Z}^N)$. Suppose that*

$$\mathrm{Vol}(\Omega) > 2^N \mathrm{Vol}(\Lambda_M) = 2^N |\det(M)|.$$

Then there exists $x \in \Omega \cap (\Lambda_M \setminus (0, \ldots, 0))$.

[2]This is the very important geometric interpretation of determinants, which we would like to assume is familiar from linear algebra. Although we have some concerns as to the validity of this assumption, we will stick with it nonetheless.

1.5. Pick's Theorem.

Following Murty and Thain, we apply Minkowski's Theorem to prove a truly classic result lying on the border of number theory and plane geometry. A **lattice polygon** is a polygon $P \subset \mathbb{R}^2$ all of whose vertices are elements of $\mathbb{Z}^2$. For the sake of definiteness, let us agree that P refers to the polygon *together with its interior*. We denote the boundary by ∂P and the interior of P by P°. And here we go!

THEOREM 13.12. *(G.A. Pick* [**Pi99**]*) Let P be a convex lattice polygon. Let*

$$I := \#(P^\circ \cap \mathbb{Z}^2) \text{ be the number of interior lattice points,}$$

$$B := \#(\partial P \cap \mathbb{Z}^2) \text{ be the number of boundary lattice points.}$$

Then we have

$$\text{Area}(P) = I + \frac{B}{2} - 1.$$

PROOF. (Murty-Thain [**MT07**])
An **elementary triangle** is a lattice triangle that has no lattice points other than its vertices. Observe that Pick's Theorem for elementary triangles states that every elementary triangle has area $\frac{1}{2}$. Our strategy of proof is as follows: we first prove Pick's Theorem for elementary triangles and then deduce the general case by an induction / triangulation argument.

Before beginning the argument, we wish to nail down a simple geometric construction, the *parallelogrammization of a triangle with respect to a vertex*. Let A, B, C be the vertices of a triangle. Let v be the vector BA and let W be the vector BC. Let D be the vector $v + w$. Then ADC is a triangle congruent to ABC and the two triangles together form the fundamental parallelogram of v and w (namely, the set of all linear combinations $\alpha v + \beta w$ with $\alpha, \beta \in [0, 1]$). If A, B and C are all lattice points, then so is D.[3]
Step 1: Let ABC be an elementary triangle. Parallelogrammizing with respect to B we get a lattice parallelogram. The area of the lattice parallelogram is the magnitude of the cross product $v \times w = (A - B) \times (C - B)$. The usual determinant formula for the cross product shows that since $v, w \in \mathbb{Z}^2$, so is $v \times w$. Every nonzero vector in $\mathbb{Z}^2$ has length at least one, so the area of the parallelogram is at least one and thus the area of the triangle ABC is at least $\frac{1}{2}$.
Step 2: Let $t = ABC$ be an elementary triangle. Seeking a contradiction, we assume that its area is greater than $\frac{1}{2}$. It is no loss of generality to assume that $B = 0$: translating with respect to a lattice point preserves elementary triangles and areas. If we parallelogramize with respect to all three vertices, we get three more elementary triangles, each congruent to the given triangle. The union of these four triangles is again a lattice triangle, say T, no longer elementary because each edge contains a lattice point that is not a vertex, but still without any interior lattice points. Let Ω be the region $T \cup -T$ obtained from T by reflecting through B (which we have assumed is 0). Then Ω is a convex body which is made up of eight congruent copies of the original elementary triangle t, so it has area greater than 4. It has no interior lattice points, and thus removing the boundary, we get a symmetric convex body Ω° of area greater than 4 and with no lattice points except the origin, contradicting Theorem 13.4!

[3]For the skeptical: let $A = (x_A, y_A)$, $B = (x_B, y_B)$, $C = (x_C, y_C)$; then $v = (x_A - x_B, y_A - y_B)$ and $w = (x_C - x_B, y_C - y_B)$, so $D = B + v + w = (x_A + x_C - x_B, y_A + y_C - y_B)$.

Step 3: Let P be any convex lattice n-polygon. We first dissect P into lattice triangles. Indeed the convexity hypothesis makes this trivial: of course we may assume $n \geq 4$; for any consecutive vertices A, B, C, drawing the diagonal AC dissects P into a lattice triangle and a convex $(n-1)$-gon (the $(n-1)$-gon is obtained by intersecting P with a half-plane with boundary line $\overleftrightarrow{AC}$, and the intersection of convex sets is convex), and now we proceed by induction. We now further dissect each lattice triangle into elementary triangles, which is also easy: if we have a lattice point on an edge that is not a vertex, connect it to the opposite side, bisecting the triangle; if we have a lattice point in the interior of the triangle, connect it to each of the vertices, trisecting the triangle.

We now prove Pick's Theorem by induction on the number of elementary triangles in the dissection; we did the base case $n = 1$ already. Assume it holds for unions of n elementary triangles, and let P_{n+1} be a convex lattice polygon that is the union of $n+1$ elementary triangles. So P_{n+1} consists of a convex lattice n-gon P_n together with one more elementary triangle t. Let us write

$$I_n \text{ for the number of interior lattice points of } P_n,$$
$$B_n \text{ for the number of boundary lattice points of } P_n,$$
$$I_{n+1} \text{ for the number of interior lattice points of } P_{n+1},$$
$$B_{n+1} \text{ for the number of boundary lattice points of } P_{n+1}.$$

By induction we have

$$\text{Area}(P_n) = I_n + \frac{B_n}{2} - 1.$$

We also have $I_{n+1} = I_n$ and $B_{n+1} = B_n + 1$, and thus

$$I_{n+1} + \frac{B_{n+1}}{2} - 1 = I_n + \frac{B_n + 1}{2} - 1 = (I_n + \frac{B_n}{2} - 1) + \frac{1}{2}$$
$$\stackrel{\text{IH}}{=} \text{Area}(P_n) + \frac{1}{2} = \text{Area}(P_n) + \text{Area}(t) = \text{Area}(P_{n+1}). \qquad \square$$

EXERCISE 13.1. Show that there are infinitely many pairwise noncongruent elementary triangles.

Regarding the convexity hypothesis in the statement of Pick's Theorem: this does not appear in all formulations. We put it there so as to make it easy to see that the polygon can be dissected into elementary triangles. In fact the theorem holds for any *simple* polygon – namely, such that the boundary is a simple closed curve – but this requires the fact that any simple polygon can be triangulated (in fact, lattice triangulated, but that is not any harder).

EXERCISE 13.2. Prove it. Namely, show that any simple lattice polygon can be dissected into elementary triangles. Deduce that Pick's Theorem holds for all *simple lattice polygons*.

Without the simplicity hypothesis, the question of precisely what one means by a polygon becomes a pertinent one, and construed sufficiently inclusively, the result can fail. As an example, consider a region between concentric squares:

$$H := [-2,2]^2 \setminus (-1,1)^2.$$

Then $\text{Area}(H) = 4^2 - 2^2 = 12$, $I = 0$ and $B = 24$, so we have

$$\text{Area}(H) = I + \frac{B}{2}$$

rather than

$$\text{Area}(H) = I + \frac{B}{2} - 1.$$

Those who know about Euler characteristics may notice that while the Euler characteristic of any simple polygon is 1, the Euler characteristic of H is 0. This suggests the following generalization of Pick's Theorem.

THEOREM 13.13. *(Extended Pick's Theorem) Let P be a closed subset of the plane whose boundary consists of finitely many lattice polygons. Let*

$$I := \#(P^\circ \cap \mathbb{Z}^2) \text{ be the number of interior lattice points,}$$

$$B := \#(\partial P \cap \mathbb{Z}^2) \text{ be the number of boundary lattice points,}$$

$$\chi(P) \text{ be the Euler characteristic of } P.$$

Then we have

$$\text{Area}(P) = I + \frac{B}{2} - \chi(P).$$

EXERCISE 13.3. Prove Theorem 13.13 (e.g. by reducing to Theorem 13.12).

EXERCISE 13.4. Let $\Lambda \subset \mathbb{R}^2$ be a lattice. Formulate a generalization of Theorem 13.12 (or, if you like, of Theorem 13.13) for convex polygons whose vertices are elements of Λ.

1.6. Comments and complements.

Theorem 13.4 was first proved in an 1896 paper of H. Minkowski, and is treated at further length in Minkowski's 1910 text *Geometrie der Zahlen* [**Mi10**, pp. 73-76]. Another proof is given in his 1927 *Diophantische Approximationen* [**Mi27**, pp. 28-30]. Theorem 13.7 appears in a 1914 paper of H.F. Blichfeldt [**Bl14**], and the connection to Minkowski's theorem is noted therein. Our first proof of Theorem 13.4 – which seems to me to be the most direct – is due to Mordell [**Mo34**].

Blichfeldt's Theorem is equivalent to the following result:

THEOREM 13.14. *Let $\Omega \subset \mathbb{R}^N$ be a bounded (Jordan or Lebesgue) measurable subset of volume greater than one. Then there exists $x \in \mathbb{R}^N$ such that the translate $x + \Omega$ contains at least two integral points.*

We leave the proof as an exercise.

There is also a "rotational analogue" of Blichfeldt's Theorem:

THEOREM 13.15. *(Hammer [**Ham68**]) Let $\Omega \subset \mathbb{R}^N$ be a convex body. If the volume of Ω is greater than the volume of the unit ball in $\mathbb{R}^N$, then there exists an orthogonal matrix $M \in O(N)$ such that $M\Omega$ contains a nonzero lattice point.*

The proof is not so hard, but it uses some further facts about convex bodies.

Minkowski's theorem is often regarded as the "fundamental theorem" upon which an entire field, the **geometry of numbers**, is based. Because of this, it is not surprising that many mathematicians – including Minkowski himself and C.L. Siegel – have given various refinements over the years. Below we describe one such refinement which can be proved along similar lines.

First, we may allow the nonempty, centrally symmetric convex set $\Omega \subset \mathbb{R}^N$ to be unbounded. In order to do this, we need to make sense of Jordan measurability and volume for an unbounded subset Ω. Since we still want to define $\mathrm{Vol}(\Omega) = \int_{\mathbb{R}^N} \mathbf{1}_\Omega$, it comes down to defining what it means for a function defined on an unbounded subset of $\mathbb{R}^N$ to be Riemann integrable. Evidently what we want is an improper multivariable Riemann integral. Recall that for improper integrals over the real line, if the function f is allowed to take both positive and negative values then we need to be extremely precise about the sense in which the limits are taken, but if f is a non-negative function all roads lead to the same answer. Note that characteristic functions are non-negative. So the following definition is simple and reasonable:

Let $f : \mathbb{R}^N \to [0, \infty)$ be a function such that the restriction of f to any rectangle $[a, b] = \prod_{i=1}^N [a_i, b_i]$ is Riemann integrable. Then we define

$$\int_{\mathbb{R}^N} f = \sup \int_{[a,b]} f,$$

where the supremum ranges all integrals over all rectangles. Note that such an improper integral is always defined although it may be ∞: for instance it will be if we integrate the constant function 1 over $\mathbb{R}^N$.

THEOREM 13.16. *(Refined Minkowski Theorem) Let $\Omega \subset \mathbb{R}^N$ be a nonempty centrally symmetric convex subset.*
a) Then $\#(\Omega \cap \mathbb{Z}^N) \geq 2(\lceil \frac{\mathrm{Vol}(\Omega)}{2^N} \rceil - 1) + 1$.
b) If Ω is closed and bounded, then $\#(\Omega \cap \mathbb{Z}^N) \geq 2(\lfloor \frac{\mathrm{Vol}(\Omega)}{2^N} \rfloor) + 1$.

In other words, part a) says that if for some positive integer k we have $\mathrm{Vol}(\Omega)$ is strictly greater than $k \cdot 2^N$, then Ω contains at least $2k$ nonzero lattice points (which necessarily come in k antipodal pairs P, $-P$). Part b) says that the same conclusion holds in the limiting case $\mathrm{Vol}(\Omega) = k \cdot 2^N$ provided Ω is closed and bounded.

There are analogous refinements of Blichfeldt's theorem; moreover, by a linear change of variables we can get a "Refined Mark II Minkowski Theorem" with the standard integral lattice $\mathbb{Z}^N$ replaced by any lattice $\Lambda = M\mathbb{Z}^N$, with a suitable correction factor of $\mathrm{Vol}(\Lambda)$ thrown in.

We leave the proof of Theorem 13.16 and the statements and proofs of these other refinements as exercises for the interested reader.

2. Diophantine Applications

2.1. The Two Squares Theorem Again.

Suppose $p = 4k + 1$ is a prime number.

By Fermat's Lemma (Lemma 2 of Handout 4), there exists $u \in \mathbb{Z}$ such that $u^2 \equiv -1 \pmod{p}$: equivalently, u has order 4 in $(\mathbb{Z}/p\mathbb{Z})^\times$. Define

$$M := \begin{bmatrix} 1 & 0 \\ u & p \end{bmatrix}.$$

We have $\det(M) = p^2$, so $\Lambda := M\mathbb{Z}^2$ defines a lattice in $\mathbb{R}^2$ with

$$\mathrm{Vol}(\Lambda) = \det(M)\,\mathrm{Vol}(\mathbb{Z}^2) = p.$$

If $(t_1, t_2) \in \mathbb{Z}^2$ and $(x_1, x_2)^t = M(t_1, t_2)^t$, then

$$x_1^2 + x_2^2 = t_1^2 + (ut_1 + pt_2)^2 \equiv (1 + u^2)t_1^2 \equiv 0 \pmod{p}.$$

Now let

$$\Omega = B_0(\sqrt{2p}) = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 2p\}$$

be the open ball of radius $2\sqrt{p}$ about the origin in $\mathbb{R}^2$. We have

$$\mathrm{Vol}\,\Omega = \pi(\sqrt{2p})^2 = 2\pi p > 4p = 2^2\,\mathrm{Vol}\,\Lambda,$$

so by Minkowski's Theorem Mark II there exists $(x_1, x_2) \in \Lambda$ with

$$0 < x_1^2 + x_2^2 < 2p.$$

Since $p \mid x_1^2 + x_2^2$, the only possible conclusion is

$$x_1^2 + x_2^2 = p.$$

2.2. The Four Squares Theorem.

LEMMA 13.17. *(Euler's Identity) For any integers $a_1, \ldots, a_4, b_1, \ldots, b_4$, we have*

$$(a_1^2 + a_2^2 + a_3^2 + a_4^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2) = (a_1b_1 - a_2b_2 - a_3b_3 - a_4b_4)^2 +$$
$$(a_1b_2 + a_2b_1 + a_3b_4 - a_4b_3)^2 + (a_1b_3 - a_2b_4 + a_3b_1 + a_4b_2)^2 + (a_1b_4 + a_2b_3 - a_3b_2 + a_4b_1)^2.$$

PROOF. Exercise! □

Thus the set of sums of four integer squares is closed under multiplication. Since $1 = 1^2 + 0^2 + 0^2 + 0^2$ is a sum of four squares, it suffices to show that each prime p is a sum of four squares. Since $2 = 1^2 + 1^2 + 0^2 + 0^2$, we may assume $p > 2$.

LEMMA 13.18. *The (four-dimensional) volume of a ball of radius r in $\mathbb{R}^4$ is $\frac{\pi^2}{2}r^4$.*

PROOF. Exercise! □

LEMMA 13.19. *For a prime $p > 2$ and $a \in \mathbb{Z}$, there exist $r, s \in \mathbb{Z}$ such that*

$$r^2 + s^2 \equiv a \pmod{p}.$$

PROOF. There are $\frac{p-1}{2}$ nonzero squares mod p and hence $\frac{p-1}{2}+1 = \frac{p+1}{2}$ squares mod p. Rewrite the congruence as $r^2 \equiv a - s^2 \pmod{p}$. Since the map $\mathbb{F}_p \to \mathbb{F}_p$ given by $t \mapsto a - t$ is an injection, as x ranges over all elements of $\mathbb{F}_p$ both the left and right hand sides take $\frac{p+1}{2}$ distinct values. Since $\frac{p+1}{2} + \frac{p+1}{2} > p$, these subsets cannot be disjoint, and any common value gives a solution to the congruence. □

THEOREM 13.20. *(Lagrange) Every positive integer is a sum of four integral squares.*

PROOF. By Lemma 13.19, there are $r, s \in \mathbb{Z}$ such that $r^2 + s^2 + 1 \equiv 0 \pmod{p}$. Define

$$M = \begin{bmatrix} p & 0 & r & s \\ 0 & p & s & -r \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We have $\det(M) = p^2$, so $\Lambda := M\mathbb{Z}^4$ defines a lattice in $\mathbb{R}^4$ with

$$\mathrm{Vol}(\Lambda) = \det(M)\,\mathrm{Vol}(\mathbb{Z}^4) = p^2.$$

If $(t_1, t_2, t_3, t_4) \in \mathbb{Z}^4$ and $(x_1, x_2, x_3, x_4) := M(t_1, t_2, t_3, t_4)$ then

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 + x_4^2 &= (pt_1 + rt_3 + st_4)^2 + (pt_2 + st_3 - rt_4)^2 + t_3^2 + t_4^2 \\ &\equiv\ t_3^2(r^2 + s^2 + 1) + t_4^2(r^2 + s^2 + 1)\ \equiv 0 \pmod{p}. \end{aligned}$$

Now let

$$\Omega = B_0(\sqrt{2p}) = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_1^2 + x_2^2 + x_3^2 + x_4^2 < 2p\}$$

be the open ball of radius $\sqrt{2p}$ about the origin in $\mathbb{R}^4$. Using Lemma 13.18 we have

$$\mathrm{Vol}(\Omega) = \frac{\pi^2}{2}(\sqrt{2p})^4 = 2\pi^2 p^2 > 16p^2 = 2^4\,\mathrm{Vol}\,\Lambda,$$

so by Minkowski's Theorem Mark II there exists $(x_1, \ldots, x_4) \in \Lambda$ with

$$0 < x_1^2 + x_2^2 + x_3^2 + x_4^2 < 2p.$$

Since $p \mid x_1^2 + x_2^2 + x_3^2 + x_4^2$, the only possible conclusion is

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = p.$$

□

2.3. Vista: Testing for a PID.

The preceding applications were very pretty, but – in that they give new proofs of old theorems – do not really serve to illustrate the power and utility of Minkowski's Convex Body Theorem. A much deeper application is to the computation of the **class number** of a number field K. Although it will be beyond us to give proofs, we feel the concept is so important that we should at least sketch the statement.

Let K be any number field, i.e., a field which contains $\mathbb{Q}$ as a subfield and is finite-dimensional as a $\mathbb{Q}$-vector space, say of dimension d. To a number field we attach its **ring of integers** $\mathbb{Z}_K$. This is the set of all elements α in K which satisfy a monic polynomial with integral coefficients: i.e., for which there exist $a_0, \ldots, a_{n-1} \in \mathbb{Z}$ such that

$$\alpha^n + a_{n-1}\alpha^{n-1} + \ldots + a_1\alpha + a_0 = 0.$$

It is not hard to show that $\mathbb{Z}_K$ is indeed a subring of K: this is shown in Handout A3. But more is true: if d is the degree of K over $\mathbb{Q}$, then there exist $\alpha_1, \ldots, \alpha_d \in \mathbb{Z}_K$ such that every element $\alpha \in \mathbb{Z}_K$ can uniquely be expressed as a $\mathbb{Z}$-linear combination of the α_i's:

$$\alpha = a_1\alpha_1 + \ldots + a_d\alpha_d,\ a_i \in \mathbb{Z}.$$

Such a d-tuple $(\alpha_1, \ldots, \alpha_d)$ of elements of $\mathbb{Z}_K$ is called an **integral basis**.

EXAMPLE 13.21. *Let $K = \mathbb{Q}$. Then $\mathbb{Z}_K = \mathbb{Z}$, and $\alpha_1 = 1$ is an integral basis.*

EXAMPLE 13.22. *Let $K/\mathbb{Q}$ be a quadratic extension, so that there exists a squarefree integer $d \neq 0, 1$ such that $K = \mathbb{Q}(\sqrt{d})$. Observe that $\sqrt{d}$, satisfying the monic polynomial $t^2 - d$, is an element of $\mathbb{Z}_K$, as is the entire subring $\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} \mid a, b \in \mathbb{Z}\}$ that it generates. With $d = -1$, this is just the ring of Gaussian integers, which is indeed the full ring of integers of $\mathbb{Q}(\sqrt{-1})$.*

In general things are more subtle: it turns out that if $d \equiv 2, 3 \pmod 4$ then $\mathbb{Z}_K = \mathbb{Z}[\sqrt{d}]$; however if $d \equiv 1 \pmod 4$ then the element $\tau_d := \frac{1+\sqrt{d}}{2}$ may not look like an algebraic integer but it satsifies the monic polynomial $t^2 + t + \frac{1-d}{4}$ (which has $\mathbb{Z}$-coefficients since $d \equiv 1 \pmod 4$) so in fact it is, and in this case $\mathbb{Z}_K = \mathbb{Z}[\tau_d] = \{a + b(\frac{1+\sqrt{d}}{2}) \mid a, b \in \mathbb{Z}\}$.

EXAMPLE 13.23. *Let $K = \mathbb{Q}(\zeta_n)$ obtained by adjoining to $\mathbb{Q}$ a primitive nth root of unity. Then it is easy to see that ζ_n is an algebraic integer, and in this case it can be shown that $\mathbb{Z}_K = \mathbb{Z}[\zeta_n]$ is the full ring of integers.*

It is rare to be able to write down an integral basis by pure thought; however, there exists an algorithm which, given any single number field K, computes an integral basis for K.

QUESTION 8. *For which number fields K is $\mathbb{Z}_K$ a principal ideal domain?*

This is absolutely one of the deepest and most fundamental number-theoretic questions because, as we have seen, in trying to solve a Diophantine equation we are often naturally led to consider arithmetic in a ring of integers $\mathbb{Z}_K$ – e.g., in studying the equation $x^2 - Dy^2 = n$ we take $K = \mathbb{Q}(\sqrt{D})$ and in studying $x^n + y^n = z^n$ we take $K = \mathbb{Q}(\zeta_n)$. If $\mathbb{Z}_K$ turns out to be a PID, we can use Euclid's Lemma, a formidable weapon. Indeed, it turns out that a common explanation of each of the classical success stories regarding these two families of equations (i.e., theorems of Fermat, Euler and others) is that the ring $\mathbb{Z}_K$ is a PID.

Gauss conjectured that there are infinitely many squarefree $d > 0$ such that the ring of integers of the real quadratic field $K = \mathbb{Q}(\sqrt{d})$ is a PID. This is still unknown; in fact, for all we can prove there are only finitely many number fields K (of any and all degrees!) such that $\mathbb{Z}_K$ is a PID. In this regard two important goals are:

(i) To give an algorithm that will decide, for any given K, whether $\mathbb{Z}_K$ is a PID;
(ii) When it isn't, to "quantify" the failure of uniqueness of factorization in $\mathbb{Z}_K$.

For this we define the concept of **class number**. If R is any integral domain, we define an equivalence relation on the set $\mathcal{I}(R)$ of nonzero ideals of R. Namely we put $I \sim J$ iff there exist nonzero elements $a, b \in R$ such that $(a)I = (b)J$. This partitions all the nonzero ideals into equivalence classes, simply called **ideal classes**.[4] The **class number** of R is indeed the number of classes of ideals. For an arbitrary domain R, the class number may well be infinite.

The point here is that there is one distinguished class of ideals: an ideal I is equivalent to the unit ideal $R = (1)$ iff it is principal. It follows that R is a PID iff its class number is equal to one. Therefore both (i) and (ii) above would be addressed if we can compute the class number of an arbitrary ring of integers $\mathbb{Z}_K$.

This is exactly what Minkowski did:

[4]In fact, the use of the term "class" in mathematics in the context of equivalence relations can be traced back to this very construction in the case of $R = \mathbb{Z}_K$ the ring of integers of an imaginary quadratic field K, which was considered by Gauss in his *Disquisitones Arithmeticae*.

THEOREM 13.24. *(Minkowski) Let K be any number field.*
a) The ideals of the ring $\mathbb{Z}_K$ of integers of K fall into finitely many equivalence classes; therefore K has a well-defined class number $h(K) < \infty$.
b) There is an explicit upper bound on $h(K)$ in terms of invariants of K which can be easily computed if an integral basis is known.
c) There is an algorithm to compute $h(K)$.

The proof is not easy; apart from the expected ingredients of more basic algebraic number theory, it also uses, crucially, Theorem 13.4.

As an example of the usefulness of the class number in "quantifying" failure of factorization even when $\mathbb{Z}_K$ is not a UFD, we note that Lamé erroneously believed he could prove FLT for all odd primes p because he assumed (implicitly, since the concept was not yet clearly understood) that $\mathbb{Z}[\zeta_p]$ was always a PID. Lamé's proof is essentially correct when the class number of $\mathbb{Q}(\zeta_p)$ is equal to one, which *is* some progress from the previous work on FLT, but unforunately this happens iff $p \leq 19$. Kummer on the other hand found a sufficient condition for FLT(p) to hold which turns out to be equivalent to: the class number of $\mathbb{Q}(\zeta_p)$ is not divisible by p. This condition, in turn, is satisfied for all $p < 200$ *except* for 37, 59, 67, 101, 103, 131, 149, and 157; and *conjecturally* for a subset of the primes of relative density $e^{\frac{-1}{2}} \approx 0.61$. Note finally that this remains conjectural to this day while FLT has been proven: the study of class numbers really is among the deepest and most difficult of arithmetic questions.

2.4. Comments and complements.

As is the case for many of the results we have presented, one of the attractions of Theorem 13.20 is its simple statement. Anyone who is inquisitive enough to wonder which integers can be written as a sum of four squares will eventually conjecture the result, but the proof is of course another matter. Apparently the first recorded statement – without proof – is in the *Arithmetica* of Diophantus of Alexandria, some time in the third century AD. Diophantus' text entered into the mathematical consciousness of Renaissance Europe through Gaspard Bachet's 1620 Latin translation of the *Arithmetica*.

Famously, Fermat was an ardent reader of Bachet's book, and he saw and claimed a proof of the Four Squares Theorem. As we have already mentioned, with one exception (FLT for $n = 4$) Fermat *never* published proofs, making the question of exactly which of his "theorems" he had actually proved a subject of perhaps eternal debate. In this case the consensus among mathematical historians seems to be skepticism that Fermat actually had a proof. In any case, the proof was still much sought after Fermat's death in 1665. Euler, one of the greatest mathematicians of the 18th century, labored for 40 years without finding a proof. Finally the theorem was proved and published in 1770 by the younger and equally great[5] Italian-French mathematician Joseph-Louis Lagrange.

[5]In quality at least. No one has ever equalled Euler for quantity, not even the famously prolific and relatively long-lived twentieth century mathematician Paul Erdös, although there are one or two living mathematicians that might eventually challenge Euler.

There are many proofs of the Four Squares Theorem. Some are "completely elementary", i.e., which neither require nor introduce any "extraneous" concepts like lattices. The most pedestrian proof begins as ours did with Euler's identity and Lemma 13.19. From this we know that it suffices to represent any prime as a sum of four squares and also that for any prime p, some positive integer multiple mp is of the form $r^2 + s^2 + 1^2$ and in particular a sum of four squares, and the general strategy is to let m_0 be the smallest integral multiple of p which is a sum of four squares and to show, through a "descent" argument, that $m_0 = 1$. Lagrange's original proof followed this strategy, and many elementary number theory texts contain such a proof, including [**HW**].

Another proof, which has the virtue of explaining the mysterious identity of Lemma 13.17 proceeds in analogy to our first proof of the two squares theorem: it works in a certain (non-commutative!) ring of **integral quaternions**. Quaternions play a vital role modern number theory, but although it is not too hard to introduce enough quaternionic theory to prove the Four Squares Theorem (again see [**HW**]), one has to dig deeper to begin to appreciate what is really going on.

Yet another proof uses the arithmetic properties of theta series; this leads to an exact formula for $r_4(n)$, the number of representations of a positive integer as a sum of four squares. In this case, to understand what is really going on involves discussion of the arithmetic theory of modular forms, which is again too rich for our blood (but we will mention that modular forms and quaternions are themselves quite closely linked!); and again Hardy and Wright manage to give a proof using only purely formal power series manipulations, which succeeds in deriving the formula for $r_4(n)$.

Regarding generalizations of Theorem 13.20, we will only mention one: a few months *before* Lagrange's proof, Edward Waring asserted that "every number is a sum of four squares, nine cubes, nineteen biquadrates [i.e., fourth powers] and so on." In other words, Waring believed that for every positive integer k there exists a number n depending only on k such that every positive integer is a sum of n non-negative kth powers. If so, we can define $g(k)$ to be the least such integer k. Evidently the Four Squares Theorem together with the observation that 7 is not a sum of three squares, give us $g(2) = 4$. That $g(k)$ actually exists for all k is by no means obvious. This was first proven by David Hilbert in 1909. We now know the exact value of $g(k)$ for all k; that $g(3) = 9$ was established relatively early on (Wieferich, 1912), but $g(4)$ was the last value to be established, by Balasubramanian in 1986: indeed $g(4) = 19$, as Waring predicted.

Of more enduring interest is the quantity $G(k)$, defined to be the least positive integer n such that every sufficiently large positive integer can be written as a sum of n non-negative kth powers: i.e., we allow finitely many exceptions. Since for all k, $8k+7$ is not even a sum of three squares modulo 8, none of these infinitely many integers are sums of three squares, so $g(2) = G(2) = 4$. On the other hand it is known that $G(3) \leq 7 < 9 = g(3)$, and it morever *suspected* that $g(3) = 4$, but this is far from being proven. Indeed only one other value of G is known.

THEOREM 13.25. *(Davenport, 1939)* $G(4) = 16$.

Getting better bounds on $G(k)$ is an active topic in contemporary number theory.

CHAPTER 14

The Chevalley-Warning Theorem

1. The Chevalley-Warning Theorem

In this chapter we shall discuss a result that was conjectured by Emil Artin in 1935 and proved shortly thereafter by Claude Chevalley. A refinement was given by Artin's student Ewald Warning, who, as the story goes, was the one whom Artin had intended to prove the theorem before Chevalley came visiting Göttingen and got Artin to say a little too much about the problem his student was working on.

One of the charms of the Chevalley-Warning theorem is that it can be stated and appreciated without much motivational preamble. So let's just jump right in.

1.1. Statement of the theorem(s).

Let $q = p^a$ be a prime power, and let $\mathbb{F}_q$ be a finite field of order q. We saw earlier in the course that there exists a finite field of each prime power cardinality.[1] For the reader who is unfamiliar with finite fields, it may be a good idea to just replace $\mathbb{F}_q$ with $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ on a first reading, and then afterwards look back and see that the assumption of an arbitrary finite field changes nothing.

THEOREM 14.1. *(Chevalley's Theorem) Let n, $d_1, \ldots, r$ be positive integers such that $d_1 + \ldots + d_r < n$. For each $1 \leq i \leq r$, let $P_i(t_1, \ldots, t_n) \in \mathbb{F}_q[t_1, \ldots, t_n]$ be a polynomial of total degree d_i with zero constant term: $P_i(0, \ldots, 0) = 0$. Then there exists $0 \neq x = (x_1, \ldots, x_n) \in \mathbb{F}_q^n$ such that*

$$P_1(x) = \ldots = P_r(x) = 0.$$

Exercise 1: Suppose we are given any system of polynomials $P_1(t), \ldots, P_r(t)$ in n variables $t_1, \ldots, t_n$ with $\sum_i \deg(P_i) < n$. Deduce from Chevalley's that if there exists at least one $x \in \mathbb{F}_q^n$ such that $P_1(x) = \ldots = P_r(x)$, then there exists $y \neq x$ such that $P_1(y) = \ldots = P_r(y)$.
(Hint: Make a change of variables to reduce to Chevalley's theorem.)

In other words, Exercise 1 asserts that a system of polynomials in n variables over $\mathbb{F}_q$ cannot have exactly one common solution, provided the sum of the degrees is less than n. Warning's theorem gives a generalization:

THEOREM 14.2. *(Warning's Theorem) Let n, $d_1, \ldots, r$ be positive integers such that $d_1 + \ldots + d_r < n$. For each $1 \leq i \leq r$, let $P_i(t_1, \ldots, t_n) \in \mathbb{F}_q[t_1, \ldots, t_n]$ be a polynomial of total degree d_i. Let*

$$Z = \#\{(x_1, \ldots, x_n) \in \mathbb{F}_q^n \mid P_1(x_1, \ldots, x_n) = \ldots = P_r(x_1, \ldots, x_n) = 0.\}$$

[1] It can be shown that any two finite fields of the same order are isomorphic – see any text on field theory – but we don't need this uniqueness statement here.

Then $Z \equiv 0 \pmod{p}$.

Arin had conjectured the following special case:

COROLLARY 14.3. *Let* $P(t_1, \ldots, t_n) \in \mathbb{F}[t_1, \ldots, t_n]$ *be a homogeneous polynomial of degree* d *in* n *variables over a finite field* $\mathbb{F}$. *If* $n > d$ *then there exists* $(0, \ldots, 0) \neq (x_1, \ldots, x_n) \in \mathbb{F}^n$ *such that* $P(x_1, \ldots, x_n) = 0$.

In in the sequel, we will refer to any of Theorem 14.1, Theorem 14.2 or Corollary 14.3 as the **Chevalley-Warning theorem**.

1.2. Applications to Quadratic Forms.

Taking $d = 1$ Corollary 14.3 asserts that any homogenous linear equation $at_1 + bt_2 = 0$ (with a and b not both 0) over a finite field has a nonzero solution. Of course linear algebra tells us that the solution set to such an equation is a one-dimensional vector space, and this holds over *any* field, infinite or otherwise. So this is a trivial case.

Already the case $d = 2$ is much more interesting. A homogeneous polynomial of degree 2 is called a **quadratic form**. For simplicity, we shall for the most part consider here only nondegenerate diagonal forms over a field F,[2] i.e.,

$$q(x) = q(x_1, \ldots, x_n) = a_1 x_1^2 + \ldots + a_n x_n^2, \ x_1, \ldots, x_n \in F, \ x_1 \cdots x_n \neq 0.$$

For some fields F, no matter how large we take n to be, we can still choose the coefficients so that $q(x) = 0$ has no nontrivial solution. For instance, consider the sum of squares form

$$q_n(x) = x_1^2 + \ldots + x_n^2.$$

This has no nontrivial solution over the real numbers, or over any subfield of $\mathbb{R}$.

PROPOSITION 14.4. *Let* F *be any field, and consider the form* $q_a(x) = x_1^2 + ax_2^2$.
a) The form $q_{2,a}$ *has a nontrivial solution iff there exists* $\alpha \in F$ *such that* $\alpha^2 = -a$.
b) Therefore if $F = \mathbb{F}_q$, $q = p^a$, *then* $q_{2,1}(x) = x_1^2 + x_2^2$ *has a nontrivial solution iff* $p = 2$, $p \equiv 1 \pmod{4}$ *or* a *is even.*

Exercise 2: Prove Proposition 14.4.

Exercise 3: a) Suppose F is a field (of characteristic different from 2) which admits a quadratic field extension $K = F(\sqrt{\alpha})$. Deduce that there exists a binary quadratic form $q_{2,a}(x)$ over F which has no nontrivial solution.
b) For any odd $q = p^a$, show that there exists a binary quadratic form q over $\mathbb{F}_q$ with only the trivial solution. Can you write one down explicitly?
c)* Show that the conclusion of part b) still holds when q is even, although in this case one has to take a nondiagonal form $q(x, y) = ax^2 + bxy + cy^2 = 0$.

According to Corollary 14.3, any quadratic form in at least three variables over a finite field has a nontrivial solution. This is quite different from the situation for $F = \mathbb{R}$ or $F = \mathbb{Q}$. And it has many useful consequences, e.g.:

[2]When the characteristic of F is not 2, one can diagonalize every quadratic form by making a linear change of variables, so no generality is lost by restricting to the diagonal case.

PROPOSITION 14.5. *Let F be a field of characteristic different from 2 in which each quadratic form in three variables has a nontrivial solution. Then, for any $a, b, c \in F^\times$, there exist $x, y \in F$ such that*

$$ax^2 + by^2 = c.$$

PROOF. In other words, we are claiming that the **inhomogeneous** equation $ax^2 + by^2 = c$ has a solution over F. To see this, we **homogenize**: introduce a third variable z and consider the equation: $ax^2 + by^2 - cz^2 = 0$. By Corollary 14.3 there are $x_0,\ y_0,\ z_0 \in K$, not all zero, such that $ax_0^2 + by_0^2 = cz_0^2$. If $z_0 \neq 0$, then we can divide through, getting

$$a\left(\frac{x_0}{z_0}\right)^2 + b\left(\frac{y_0}{z_0}\right)^2 = c.$$

If $z_0 = 0$, this doesn't work; rather, we get a nontrivial solution (x_0, y_0) to $ax_0^2 + by_0^2 = 0$. Dividing by a we get $x_0^2 + (\frac{b}{a})y_0^2 = 0$. But as above this can only happen if $\frac{-b}{a} = t^2$ is a square in K, and then we can factor $q(x) = x^2 - t^2y^2 = (x+ty)(x-ty)$. This gives us a lot more leeway in solving the equation. For instance, we could factor c as $c \cdot 1$ and give ourselves the linear system

$$x + ty = c$$

$$x - ty = 1$$

which has a solution $(x, y) = (\frac{c+1}{2}, \frac{c-1}{2})$. Note that it is here that we use the hypothesis that the characteristic of K is not 2. □

In particular this gives an alternate (much more sophisticated!) proof of [Minkowski's Theorem Handout, Lemma 15].

2. Two proofs of Warning's theorem

2.1. Polynomials and polynomial functions.

We begin with a discussion about polynomials as ring elements versus polynomials as functions which is of interest in its own right. (In fact, it is because of the interest of these auxiliary results that we have chosen to include this proof.)

Let R be an integral domain and $R[t_1, \ldots, t_n]$ be the polynomial ring in n indeterminates over R. An element $P(t) = P(t_1, \ldots, t_n) \in R[t_1, \ldots, t_n]$ is a purely formal object: it is a finite R-linear combination of monomial terms, which are added and multiplied according to simple formal rules.

Note that this is not the perspective on polynomials one encounters in calculus and analysis. For instance a univariate polynomial $P(t) = a_nt^n + \ldots + a_1t + a_0 \in \mathbb{R}[t]$ is regarded as a **function** from $\mathbb{R}$ to $\mathbb{R}$, given of course by $x \mapsto P(x)$. Similarly for multivariable polynomials: $P(t_1, \ldots, t_n) \in \mathbb{R}[t_1, \ldots, t_n]$ may be defined by the same formal $\mathbb{R}$-linear combination of monomial terms as above but that is just notation: what matters is the function $\mathbb{R}^n \to \mathbb{R}$ given by $(x_1, \ldots, x_n) \mapsto P(x_1, \ldots, x_n)$. In other words, a polynomial in n variables can be **evaluated** at any point of $\mathbb{R}^n$.

Can these two perspectives be reconciled? A moment's thought makes it clear

that the "evaluation" of a polynomial is a perfectly algebraic operation: in other words, given any domain R and element $P(t)$ of the polynomial ring $R[t_1, \ldots, t_n]$, we can evaluate P at any point $(x_1, \ldots, x_n) \in R^n$, getting an element $P(x_1, \ldots, x_n)$. To be formal about it, we have an **evaluation map**:

$$\Phi : R[t_1, \ldots, t_n] \mapsto \mathrm{Map}(R^n, R),$$

where by $\mathrm{Map}(R^n, R)$ we just mean the set of all functions $f : R^n \to R$. In fact this map Φ has some nice algebraic structure. The set $\mathrm{Map}(R^n, R)$ of all functions from R^n to R can be made into a commutative ring in which addition and multiplication are just defined "pointwise":

$$(f + g)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) + g(x_1, \ldots, x_n),$$

$$(fg)(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \cdot g(x_1, \ldots, x_n).$$

It is straightforward to see that the evaluation map Φ is then a homomorphism of rings. Let us put

$$\mathcal{P}_n := \Phi(R[t_1, \ldots, t_n]) \subset \mathrm{Map}(R^n, R),$$

so that $\mathcal{P}_n$ is the ring of polynomial functions in n variables on R.

We are interested in the following question: if $P(t), Q(t) \in R[t_1, \ldots, t_n]$ are such that for all $(x_1, \ldots, x_n) \in R^n$ we have $P(x_1, \ldots, x_n) = P(x_1, \ldots, x_n)$ – so that P and Q give the same function from R^n to R – must $P(t) = Q(t)$ as elements of $R[t_1, \ldots, t_n]$? In other words, is Φ injective?

I hope you know that in the familiar case of $n = 1$, $R = \mathbb{R}$ the answer is "yes": two real univariate polynomials which give the same function are term-by-term equal. The proof is as follows: define $R(t) := P(t) - Q(t)$. We are given that $R(x) = 0$ for all $x \in R$. But if $R(x)$ were not the zero polynomial, it would have some degree $d \geq 0$ and basic (high school!) algebra shows that a polynomial over a field of degree d cannot have more than d roots. But $R(x)$ has infinitely many roots, so it must be the identically zero polynomial.

Evidently this argument works for univariate polynomials over any infinite field. The following is a stronger result:

PROPOSITION 14.6. *Let R be an infinite integral domain and $n \in \mathbb{Z}^+$. Then the evaluation map*

$$\Phi : R[t_1, \ldots, t_n] \to \mathcal{P}_n \subset \mathrm{Map}(R^n, R)$$

is a homomorphism of rings.
a) Moreover Φ is injective.
b) However, Φ is not surjective: not every function $f : R^n \to R$ is given by a polynomial.

PROOF. a) Again it suffices to show that if $\Phi(P(t)) = 0$, then $P(t)$ is the zero polynomial. If $n = 1$, we just showed this when R was a field. But that argument easily carries over, since every integral domain R can be embedded into a field F (namely its field of fractions). If there existed a nonzero polynomial $P(t) \in R[t]$ such that there were infinitely $x \in R$ such that $P(x) = 0$, then since $R \subset F$, there are also infinitely many $x \in F$ such that $P(x) = 0$, contradiction. Assume now

that $n > 1$. In general the theory of polynomials of several variables is signficantly harder than that of univariate polynomials, but here we can use a dirty trick:

$$R[t_1, \ldots, t_{n-1}, t_n] = R[t_1, \ldots, t_{n-1}][t_n].$$

In other words, a polynomial $P(t_1, \ldots, t_n)$ in n variables over the integral domain R may be viewed as a polynomial $Q(t_n) := (P(t_1, \ldots, t_{n-1}))(t_n)$ in one variable over the integral domain $R_{n-1} := R[t_1, \ldots, t_{n-1}]$. If $P(x_1, \ldots, x_n) = 0$ for all $(x_1, \ldots, x_n) \in R^n$ then, since $R \subset R_{n-1}$, the univariate polynomial $Q(t_n)$ has infinitely many roots in R_{n-1} and thus is identically zero by the above argument.

As for part b), for instance the function $\mathbf{1}_0 : R^n \to R$ which maps $0 \in R^n$ to 1 and every other element of R^n to 0 is not a polynomial function. You are asked to show this in Exercise 4 below. Another argument is by counting: for infinite R, the cardinality of $R[t_1, \ldots, t_n]$ is equal to the cardinality of R, whereas the total number of functions from R^n to R has cardinality $|R|^{|R^n|} = 2^{|R|} > |R|$, so "most" functions are not polynomials. □

Exercise 4: Let R be an infinite integral domain and $n \in \mathbb{Z}^+$. Show that the characteristic function 1_0 of the origin – i.e., the function which maps $0 = (0, \ldots, 0)$ to 1 and every other element of R^n to zero – is not a polynomial function. (Hint: restrict the function 1_0 to a line passing through the origin, and thereby reduce to the case $n = 1$.)

We shall not need Proposition 14.6 later on, but it is interesting to contrast the infinite case with the finite case. First of all:

LEMMA 14.7. *Let $R = \mathbb{F}_q$ be a finite integral domain (necessarily a field). Then every function $f : \mathbb{F}_q^n \to \mathbb{F}_q$ is given by a polynomial.*

PROOF. We first express the function $\mathbf{1}_0$, which takes 0 to 1 and every other element to 0, as a polynomial. Indeed, since $x^{q-1} = 1$ for $x \in \mathbb{F}_q^\times$ and $0^{q-1} = 0$, we have for all $x = (x_1, \ldots, x_n) \in \mathbb{F}_q^n$ that

$$\mathbf{1_0(x)} = \prod_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{1} - \mathbf{x_i^{q-1}}).$$

For an arbitrary function $f : \mathbb{F}_q^n \to \mathbb{F}$, define

$$P_f(t) := \sum_{y \in \mathbb{F}_q^n} f(y) \prod_{i=1}^{n} (1 - (t_i - y_i)^{q-1}). \tag{43}$$

Every term in the sum with $y \neq x$ yields zero, and the term $y = x$ yields $f(x)$. □

On the other hand, over a finite field $\mathbb{F}_q$, a nonzero polynomial may evaluate to the zero function: indeed $t^q - t$ is a basic one variable example. There is no contradiction here because a nonzero polynomial over a domain cannot have more roots than its degree, but $t^q - t = \prod_{a \in \mathbb{F}_q} (t - a)$ has exactly as many roots as its degree. Moreover, no nonzero polynomial of degree less than q could lie in the kernel of the evaluation map, so $t^q - t$ is a minimal degree nonzero element of $\mathrm{Ker}(\Phi)$. But, since $\mathbb{F}_q[t]$ is a PID, every nonzero ideal I is generated by its unique monic element of least degree, so $\mathrm{Ker}(\Phi) = \langle t^q - t \rangle$.

We would like to compute $\mathrm{Ker}(\Phi)$ in the multivariable case. Reasoning as above it

is clear that for all $1 \leq i \leq n$ the polynomials $t_i^q - t_i$ must lie in the kernel of the evaluation map, so at least we have $J = \langle t_1^q - t_1, \ldots, t_n^q - t_n \rangle \subset \mathrm{Ker}(\Phi)$. We will see that in fact $J = \mathrm{Ker}(\Phi)$. We can even do better: for each polynomial $P(t)$ we can find a canonical element $\tilde{P}$ of the coset $P(t) + \mathrm{Ker}(\Phi)$.

The key idea is that of a reduced polynomial. We say that a monomial $ct_1^{a_1} \cdots t_n^{a_n}$ is **reduced** if $a_i < q$ for all i. A polynomial $P \in \mathbb{F}_q[t]$ is **reduced** if each of its nonzero monomial terms is reduced. Equivalently, a reduced polynomial is one for which the total degree in each variable is less than q.

Example: The polynomial $P_f(t)$ above is a sum of polynomials each having degree $q-1$ in each variable, so is reduced.

Exercise 5: The reduced polynomials form an $\mathbb{F}_q$-subspace of $\mathbb{F}_q[t_1, \ldots, t_n]$, with a basis being given by the reduced monomials.

The idea behind the definition is that if in a monomial term we had an exponent $t_i^{a_i}$ with $a_i \geq q$, then from the perspective of the associated function this is just wasteful: we have

$$x_i^{a_i} = x_i^{q+(a_i-q)} = x_i^q x_i^{a_i-q} = x_i x_i^{a_i-q} = x_i^{a_i-(q-1)}.$$

Thus by a sequence of "elementary reductions" of this type we can convert any polynomial P into a reduced polynomial $\tilde{P}$. Moreover, a little reflection makes clear that $P - \tilde{P} \in J$.

Is it possible for a given polynomial P to be congruent modulo $\mathrm{Ker}(\Phi)$ to more than one reduced polynomial? Well, the reduced polynomials form an $\mathbb{F}_q$-vector subspace of the space of all polynomials with basis given by the reduced monomials, of which there are q^n, so the total number of reduced polynomials is q^{q^n}. In fact this is also the total number of functions from $\mathbb{F}_q^n$ to $\mathbb{F}_q$. Since we know that every function is given by some reduced polynomial, it must be that evaluation map restricted to reduced polynomials is a bijection. Finally, since we showed that every polynomial was equivalent modulo J to a reduced polynomial, so that $\#\mathbb{F}_q[t]/J \leq q^{q^n}$. By surjectivity of Φ we know $\#\mathbb{F}_q[t]/\mathrm{Ker}(\Phi) = \#\,\mathrm{Map}(\mathbb{F}_q^n, \mathbb{F}_q) = q^{q^n}$. Therefore the quotient map $\mathbb{F}_q[t]/J \to \mathbb{F}_q[t]/\mathrm{Ker}(\Phi)$ is a bijection and hence $J = \mathrm{Ker}(\Phi)$.

Remark: More standard is to prove that a nonzero reduced polynomial does not induce the zero function by induction on the number of variables. Then the surjectivity of Φ can be deduced from the injectivity on reduced polynomials by noticing, as we did, that the domain and codomain are finite sets with the same cardinality. Our treatment here is undeniably more complicated than this, but also seems more interesting. It will also smooth the way for our first proof of Warning's theorem.

Let us summarize all the preceding results:

THEOREM 14.8. *(Polynomial evaluation theorem) Let R be an integral domain and $n \in \mathbb{Z}^+$. Let $\Phi : R[t] = R[t_1, \ldots, t_n] \to \mathrm{Map}(R^n, R)$ be the homomorphism of rings obtained by associating to each polynomial the corresponding polynomial function $x = (x_1, \ldots, x_n) \mapsto P(x)$.*

a) If R is infinite, then Φ is injective but not surjective: every function $f : R^n \to R$ is represented by at most one polynomial, and there exist functions not represented by any polynomial.
*b) If R is finite, then Φ is surjective but not injective: its kernel is the ideal $\langle t_1^q - t_1, \ldots, t_n^q - t_n \rangle$. Thus every function $f : R^n \to R$ is represented by infinitely many polynomials. Moreover, for each f there exists a unique **reduced** polynomial representative, given explicitly as the polynomial $P_f(t)$ of (43) above.*

If $f : \mathbb{F}_q^n \to \mathbb{F}_q$ is any function, we define its **reduced degree** in t_i to be the degree in t_i of the associated reduced polynomial, and similarly its **reduced total degree** to be the total degree of the associated reduced polynomial.

Exercise 6: Show that if P is any polynomial, the total degree $\deg(\tilde{P})$ of $\tilde{P}$ is less than or equal to the total degree $\deg(P)$ of P.

2.2. First proof of Warning's Theorem.

We have polynomials $P_1(t), \ldots, P_r(t)$ in n variables with $\sum_{i=1}^r \deg(P_i) < n$. Put

$$Z = \{(x_1, \ldots, x_n) \in \mathbb{F}_q^n \mid P_1(x) = \ldots = P_r(x) = 0.\} \tag{44}$$

We want to show that $\#Z \equiv 0 \pmod{p}$. Let $\mathbf{1}_Z : \mathbb{F}_q^n \to \mathbb{F}_q$ be the ($\mathbb{F}_q$-valued) "characteristic function" of the subset Z, i.e., the function which maps x to 1 if $x \in Z$ and x to 0 otherwise. Now one polynomial representative $\mathbf{1}_Z$ is

$$P(t) := \prod_{i=1}^r \left(1 - P_i(t)^{q-1}\right); \tag{45}$$

whereas – essentially by (43) above – the reduced polynomial representative is

$$Q_Z(t) = \sum_{x \in Z} \prod_{i=1}^n \left(1 - (t_i - x_i)^{q-1}\right).$$

Now comes the devilry: the total degree of $P(t)$ is $(q-1)\sum_i d_i < (q-1)n$.

On the other hand, consider the coefficient of the monomial $t_1^{q-1} \cdots t_n^{q-1}$ in $Q_Z(t)$: it is $(-1)^n \#Z$. If we **assume** that $\#Z$ is not divisible by p, then this term is nonzero and $Q_Z(t)$ has total degree at least $(q-1)n$. By Exercise X.X, we have

$$\deg(\tilde{P}) \leq \deg(P) < (q-1)n \leq \deg(Q_Z).$$

Therefore $\tilde{P} \neq \deg(Q_Z)$, whereas we ought to have $\tilde{P} = Q_Z$, since each is the reduced polynomial representative of $\mathbf{1}_Z$. Evidently we assumed something we shouldn't have: rather, we must have $p \mid \#Z$, qed.

2.3. Ax's proof of Warning's theorem.

The following BOOK PROOF of Warning's Theorem is due to James Ax [**Ax64**].

We maintain the notation of the previous section, especially the polynomial $P(t)$ of (45) and the subset Z of (44). Because $P(x) = \mathbf{1}_Z(x)$ for all $x \in \mathbb{F}_q^n$, we have

$$\#Z \equiv \sum_{x \in \mathbb{F}_q^n} P(x) \pmod{p}.$$

So we just need to evaluate the sum. Since every polynomial is an $\mathbb{F}_q$-linear combination of monomial terms, it is reasonable to start by looking for a formula for $\sum_{x \in \mathbb{F}_q^n} x_1^{a_1} \cdots x_n^{a_n}$ for non-negative integers $a_1, \ldots, a_n$. It is often the case that if $f : G \to \mathbb{C}$ is a "nice" function from an abelian group to the complex numbers, then the complete sum $\sum_{x \in G} f(x)$ has a simple expression. Case in point:

LEMMA 14.9. *Let $a_1, \ldots, a_n$ be non-negative integers.*
a) If for $1 \leq i \leq n$, a_i is a positive multiple of $q-1$, then $\sum_{x \in \mathbb{F}_q^n} x_1^{a_1} \cdots x_n^{a_n} = (-1)^n$.
b) In every other case – i.e., for at least one i, $1 \leq i \leq n$, a_i is not a positive integer multiple of $q-1$ – we have $\sum_{x \in \mathbb{F}_q^n} x_1^{a_1} \cdots x_n^{a_n} = 0$.

PROOF. Part a) is not needed in the sequel and is just stated for completeness; we leave the proof as an exercise.

As for part b), we have

$$\sum_{x \in \mathbb{F}_q^n} x_1^{a_1} \cdots x_n^{a_n} = \prod_{i=1}^n \left(\sum_{x_i \in \mathbb{F}_q} x_i^{a_i} \right).$$

By assumption there exists at least one i, $1 \leq i \leq n$, such that a_i is either 0 or is positive but not a multiple of $q-1$. If $a_i = 0$, then $\sum_{x_i \in \mathbb{F}_q} x_i^{a_i} = \sum_{x_i \in \mathbb{F}_q} 1 = q \equiv 0 \in \mathbb{F}_q$, so assume that a_i is positive but not divisible by $q-1$. Let α be a generator for the cyclic group $\mathbb{F}_q^\times$, and put $\beta = \alpha^{a_i}$. Then

$$\sum_{x_i \in \mathbb{F}_q} x_i^{a_i} = 0^{a_i} + \sum_{x_i \in \mathbb{F}_q^\times} x_i^{a_i} = 0 + \sum_{N=0}^{q-2} (\alpha^N)^{a_i} = \sum_{N=0}^{q-2} \beta^N = \frac{1-\beta^{q-1}}{1-\beta} = \frac{1-1}{1-\beta} = 0.$$

□

Finally, the polynomial $P(t)$ has degree $\sum_{i=1}^r d_i(q-1) = (q-1)\sum_{i=1}^r d_i < (q-1)n$. Thus in each monomial term $ct_1^{a_1} \cdots t_n^{a_r}$ in $P(t)$ must have $a_1 + \ldots + a_r < (q-1)n$, so it can't be the case that each $a_i \geq q-1$. Therefore Lemma 14.9 applies, and $\sum_{x \in \mathbb{F}_q^n}$ is an $\mathbb{F}_q$-linear combination of sums each of which evaluates to 0 in $\mathbb{F}_q$ and therefore the entire sum is 0. This completes our second proof of Warning's Theorem.

3. Some Later Work

Under the hypotheses of Warning's theorem we can certainly have 0 solutions. For instance, we could take $P_1(t)$ to be any polynomial with $\deg(P_1) < \frac{n}{2}$ and $P_2(t) = P_1(t) + 1$. Or, when q is odd, let $a \in \mathbb{F}_q$ be a quadratic nonresidue, let $P_1(t)$ be a polynomial of degree less than $\frac{n}{2}$ and put $P(t) = P_1(t)^2 - a$.

On the other hand, it is natural to wonder: in Warning's theorem, we might actually have $\#Z \equiv 0 \mod q$? The answer is now known, but it took 46 years.

First consider the case of $r = 1$, i.e., a single polynomial P of degree less than n. In [**Wa36**], Warning proved that $\#Z$, if positive, is at least q^{n-d}. And in the same paper [**Ax64**], Ax showed that $q^b \mid \#Z$ for all $b < \frac{n}{d}$. By hypothesis we can take $b = 1$, so the aforementioned question has an affirmative answer in this case.

For the case of multiple polynomials $P_1, \ldots, P_r$ of degrees $d_1, \ldots, d_r$, in a celebrated 1971 paper N. Katz showed that $q^b \mid \#Z$ for all positive integers b satisfying

$$b < \frac{n - (d_1 + \ldots + d_r)}{d_1} + 1.$$

Since the above fraction is by hypothesis strictly positive, we can take $b = 1$ getting indeed $\#Z \equiv 0 \pmod{q}$ in all cases.

These divisibilities are called estimates of **Ax-Katz type**. It is known that there are examples in which the Ax-Katz divisibilities are best possible, but refining these estimates in various cases is a topic of active research: for instance there is a 2007 paper by W. Cao and Q. Sun, *Improvements upon the Chevalley-Warning-Ax-Katz-type estimates*, J. Number Theory 122 (2007), no. 1, 135–141.

Notice that the work since Warning has focused on the problem of getting best possible p-adic estimates for the number of solutions: that is, instead of bounds of the form $\#\mathbb{Z} \geq N$, we look for bounds of the form $\mathrm{ord}_p(\#Z) \geq N$. Such estimates are closely linked to the **p-adic cohomology** of algebraic varieties, a beautiful (if technically difficult) field founded by Pierre Deligne in his landmark paper "Weil II."

The hypotheses of the Chevalley-Warning theorem are also immediately suggestive to algebraic geometers: (quite) roughly speaking there is a geometric division of algebraic varieties into three classes: Fano, Calabi-Yau, and general type. The degree conditions in Warning's theorem are precisely those which give, among the class of algebraic varieties represented nicely by r equations in n variables ("smooth complete intersections"), the Fano varieties. A recent result of Hélène Esnault gives the geometrically natural generalization: any Fano variety over $\mathbb{F}_q$ has a rational point. There are similar results for other Fano-like varieties.

CHAPTER 15

Additive Combinatorics

1. The Erdős-Ginzburg-Ziv Theorem

1.1. A Mathematical Card Game.

Consider the following game. One starts with a deck of one hundred cards (or N cards, for some arbitrary positive integer N). Any number of players may play; one of them is the dealer. The dealer shuffles the deck, and the player to the dealer's left selects a card ("any card") from the deck and shows it to everyone. The player to the dealer's right writes down the numerical value of the card, say n, and keeps this in a place where everyone can see it. The card numbered n is reinserted into the deck, which is reshuffled. The dealer then deals cards face up on the table, one at a time, at one minute intervals, or sooner by unanimous consent (i.e., if everyone wants the next card, including the dealer, then it is dealt; otherwise the dealer waits for a full minute). A player wins this round of the game by correctly selecting any $k > 0$ of the cards on the table such that the sum of their numerical values is divisible by n. When all the cards are dealt, the players have as much time as they wish.

For example, suppose that $n = 5$ and the first card dealt is 16. 16 is not divisible by 5, so the players all immediately ask for another card: suppose it is 92. 92 is not divisible by 5 and neither is $92 + 16 = 118$, so if the players are good, they will swiftly ask for the next card. Suppose the next card is 64. Then someone can win by collecting the 64 and the 16 and calling attention to the fact that $64 + 16 = 80$ is divisible by 5.

Here's the question: is it always possible to win the game, or can all the cards be dealt with no solution?

We claim that it is never necessary to deal more than n cards before a solution exists. Moreover, if the number N of cards in the deck is sufficiently large compared to the selected modulus n, it is possible for fewer than n cards to be insufficient.

To see the latter, note that if $n = 1$ we obviously need n cards, and if $n = 2$ we will need n cards iff the first card dealt is odd. If $n = 3$ we may need n cards iff $N \geq 4$, since if 1 and 4 are the first two cards dealt there is no solution. In general, if the cards dealt are $1, 1+n, 1+2n, \ldots, 1+(n-2)n$, then these are $n-1$ cards which are all $1 \pmod{n}$ and clearly we cannot obtain $0 \pmod{n}$ by adding up the values of any $0 < k \leq n-1$ of these. This is possible provided $N \geq n^2-2n+1 = (n-1)^2$.[1]

[1]We neglect the issue of figuring out exactly how many card are necessary if n is moderately large compared to N. It seems interesting but does not segue into our ultimate goal.

But why are n cards always sufficient? We can give an explicit algorithm for finding a solution: for each $1 \leq k \leq n$, let $S_k = a_1 + \ldots + a_k$ be the sum of the values of the first k cards. If for some k, S_k is divisible by n, we are done: we can at some point select all the cards. Otherwise, we have a sequence $S_1, \ldots, S_n$ of elements in $\mathbb{Z}/n\mathbb{Z}$, none of which are $0 \pmod{n}$. By the pigeonhole principle, there must exist $k_1 < k_2$ such that $S_{k_1} \equiv S_{k_2} \pmod{n}$, and therefore

$$0 \equiv S_{k_2} - S_{k_1} = a_{k_1+1} + \ldots + a_{k_2} \pmod{n}.$$

In other words, not only does a solution exist, for some $k \leq n$ a solution exists which we can scoop up quite efficiently, by picking up a consecutive run of cards from right to left starting with the rightmost card.

Notice that this is not always the only way to win the game, so if this is the only pattern you look for you will often lose to more skillful players. For instance, in our example of $n = 5$, the sequence (which we will now reduce mod 5) $1, 2, 4$ already has a solution but no consecutively numbered solution.

An interesting question that we will leave the reader with is the following: fix n and assume that N is much larger than n: this is effectively the same as drawing with replacement (because after we a draw any one card a_i, the change in the proportion of the cards in the deck which are congruent to $a_i \pmod{n}$ is negligible if N is sufficiently large, and we will never deal more than n cards). Suppose then that we deal $1 \leq k \leq n$ cards. What is the probability that a solution exists?

Anyway, we have proven the following amusing mathematical fact:

THEOREM 15.1. *Let $a_1, \ldots, a_n$ be any integers. There exists a nonempty subset $I \subset \{1, \ldots, n\}$ such that $\sum_{i \in I} a_i \equiv 0 \pmod{n}$.*

1.2. The Erdős-Ginzburg-Ziv Theorem.

After a while it is tempting to change the rules of any game. Suppose we "make things more interesting" by imposing the following additional requirement: we deal cards in sequence as before with a predetermined "modulus" $n \in \mathbb{Z}^+$. But this time, instead of winning by picking up any (positive!) number of cards which sum to 0 modulo n, we must select precisely n cards $a_{i_1}, \ldots, a_{i_n}$ such that $a_{i_1} + \ldots + a_{i_n} \equiv 0 \pmod{n}$. Now (again assuming that $N \gg n$, or equivalently, dealing with replacement), is it always possible to win eventually? If so, how many cards must be dealt?

Well, certainly at least n: since the problem is more stringent than before, again if the first $n-1$ congruence classes are all $1 \pmod{n}$ then no solution exists. If we have at least n instances of 1 mod n then we can take them and win. On the other hand, if the first $n-1$ cards are all 1's, then by adding up any $k \leq n-1$ of them we will get something strictly less than n, so if the next few cards all come out to be $0 \pmod{n}$, then we will not be able to succeed either. More precisely, if in the first $2n-2$ cards we get $n-1$ instances of $1 \pmod{n}$ and $n-1$ instances of $0 \pmod{n}$, then there is no way to select precisely n of them that add up to $0 \pmod{n}$. Thus at least $2n-1$ cards may be required. Conversely:

THEOREM 15.2. *(Erdős-Ginzburg-Ziv, 1961) Let $n \in \mathbb{Z}^+$ and $a_1, \ldots, a_{2n-1} \in \mathbb{Z}$. There exists a subset $I \subset \{1, \ldots, 2n-1\}$ such that:*
(i) $\#I = n$.
(ii) $\sum_{i \in I} a_i \equiv 0 \pmod{n}$.

PROOF. (C. Bailey and R.B. Richter) The first step is to deduce the theorem for $n = p$ a prime using Chevalley-Warning. The second step is to show that if the theorem holds for n_1 and for n_2, it holds also for $n_1 n_2$.

Step 1: Suppose $n = p$ is a prime number. Let $a_1, \ldots, a_{2p-1} \in \mathbb{Z}$. Consider the following elements of the polynomial ring $\mathbb{F}_p[t_1, \ldots, t_{2p-1}]$:

$$P_1(t_1, \ldots, t_{2p-1}) = \sum_{i=1}^{2p-1} a_i t_i^{p-1},$$

$$P_2(t_1, \ldots, t_{2p-1}) = \sum_{i=1}^{2p-1} t_i^{p-1}.$$

Since $P_1(0) = P_2(0) = 0$ and $\deg(P_1) + \deg(P_2) = 2p - 2 < 2p - 1$, by Chevalley-Warning there exists $0 \neq x = (x_1, \ldots, x_{2p-1}) \in \mathbb{F}_p^{2p-1}$ such that

$$\sum_{i=1}^{2p-1} a_i x_i^{p-1} = 0, \tag{46}$$

$$\sum_{i=1}^{2p-1} x_i^{p-1} = 0. \tag{47}$$

Put

$$I = \{1 \leq i \leq 2p-1 \mid x_i \neq 0\}.$$

Since (as usual!) x^{p-1} is equal to 1 if $x \neq 0$ and 0 if $x = 0$, (46) and (47) yield:

$$\sum_{i \in I} a_i \equiv 0 \pmod{p},$$

$$\sum_{i \in I} 1 \equiv 0 \pmod{p}.$$

But we have $0 < \#I < 2p$, and therefore $\#I = p$, completing the proof of Step 1.

Step 2: Because we know the theorem is true for all primes n, by induction we may assume that $n = km$ for $1 < k,\ m < n$ (i.e., n is composite) and, by induction, that the theorem holds for k and m.

By an easy induction on r, one sees that if for any $r \geq 2$ we have $rk - 1$ integers $a_1, \ldots, a_{rk-1}$, then there are $r - 1$ pairwise disjoint subsets of $I_1, \ldots, I_{r-1}$ of $\{1, \ldots, rk-1\}$, each of size k, such that for all $1 \leq j \leq r-1$ we have $\sum_{l \in I_j} a_i \equiv 0 \pmod{k}$. Apply this with $r = 2m$ to our given set of $2n - 1 = (2mk) - 1$ integers: this gives $2m - 1$ pairwise disjoint subsets $I_1, \ldots, I_{2m-1} \subset \{1, \ldots, 2n-1\}$, each of size k, such that for all $1 \leq j \leq 2m-1$ we have

$$\sum_{i \in I_j} a_i \equiv 0 \pmod{k}.$$

Now, for each j as above, put

$$b_j = \sum_{i \in I_j} a_i, \quad b'_j = \frac{b_j}{k}.$$

We thus have $2m-1$ integers $b'_1, \ldots, b'_{2m-1}$. Again using our inductive hypothesis, there exists $J \subset \{1, \ldots, 2m-1\}$ such that $\#J = m$ and $\sum_{j \in J} b'_j \equiv 0 \pmod{m}$. Let $I = \bigcup_j I_j$. Then $\#I = km = n$ and

$$\sum_{i \in I} a_i \equiv \sum_{j \in J} \sum_{i \in I_j} a_i \equiv \sum_{j \in J} k b'_j \equiv 0 \pmod{km}.$$

□

1.3. EGZ theorems in finite groups.

This application of Chevalley-Warning – one which makes good use of our ability to choose multiple polynomials – is apparently well-known to combinatorial number theorists. But I didn't know about it until Patrick Corn brought it to my attention.

As with the Chevalley-Warning theorem itself, the EGZ theorem is sort of a prototype for a whole class of problems in combinatorial algebra. In any group G (which, somewhat unusually, we will write additively even if it is not commutative) a **zero sum sequence** is a finite sequence $x_1, \ldots, x_n$ of elements of G such that (guess what?) $x_1 + \ldots + x_n = 0$. By a **zero sum subsequence** we shall mean the sequence $x_{i_1}, \ldots, x_{i_k}$ associated to a nonempty subset $I \subset \{1, \ldots, n\}$. In this language, our Theorem 15.1 says that any sequence of n elements in $\mathbb{Z}/n\mathbb{Z}$ has a zero sum subsequence. The same argument proves the following result:

THEOREM 15.3. *Let G be a finite group (not necessarily commutative), of order n. Then any sequence $x_1, \ldots, x_n$ in G has a zero sum subsequence.*

Some EGZ-type theorems in this context are collected in the following result.

THEOREM 15.4. *(EGZ for finite groups)*
a) (Erdős-Ginzburg-Ziv, 1961) Let G be a finite solvable group of order n and $x_1, \ldots, x_{2n-1} \in G$. Then there exist distinct indices $i_1, \ldots, i_n$ (not necessarily in increasing order) such that $x_{i_1} + \ldots + x_{i_n} = 0$.
b) [**Ol76**] *Same as part a) but for any finite group.*
c) [**Su99**] *Same as part a) but the indices can be chosen in increasing order: $i_1 < \ldots < i_n$.*
d) [**Su99**] *The conclusion of part c) holds for a finite group G provided it holds for all of its Jordan-Hölder factors.*

We draw the reader's attention to the distinction between the results of parts a) and b) and those of c) and d): in the first two parts, we are allowed to reorder the terms of the subsequence, whereas in the latter two we are not. In a commutative group it makes no difference – thus, the generalization to all finite abelian groups is already contained in the original paper of EGZ – but in a noncommutative group the desire to preserve the order makes the problem significantly harder.

The inductive argument in Step 2 of Theorem 15.2 is common to all the proofs, and is most cleanly expressed in Sury's paper as the fact that the class of finite

groups for which EGZ holds is closed under extensions. Thus the case in which G is cyclic of prime order is seen to be crucial. In 1961 Erdős, Ginzburg and Ziv gave an "elementary" proof avoiding Chevalley-Warning. Nowadays there are several proofs available; a 1993 paper of Alon and Dubiner presented at Erdős' 80th birthday conference gives five different proofs. Olson's proof also uses only elementary group theory, but is not easy. In contrast, Sury's paper makes full use of Chevalley-Warning and is the simplest to read: it is only three pages long.

Sury's result has the intriguing implication that it would suffice to prove the EGZ theorem for all finite simple groups (which are now completely classified...). To my knowledge no one has followed up on this.

There is another possible generalization of the EGZ theorem to finite abelian, but non-cyclic, groups. Consider for instance $G(n,2) := Z_n \times Z_n$, which of course has order n^2. Rather than asking for the maximal length of a sequence without an n^2-term zero sum subsequence, one might ask for the maximal length of a sequence without an n-term zero sum subsequence. (One might ask many other such questions, of course, but in some sense this is the most reasonable "vector-valued analogue" of the EGZ situation.) A bit of thought shows that the analogous lower bound is given by the sequence consisting of $n-1$ instances each of $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$: in other words, this is the "obvious" sequence with no n-term zero-sum subsequence, of length $4(n-1)$. It was conjectured by A. Kemnitz in 1983 that indeed any sequence in $G(n,2)$ of length at least $4n-3$ has an n-term zero sum subsequence. Kemnitz's conjecture was proved in 2003 independently by Christian Reiher [**Re07**] (an undergraduate!) and Carlos di Fiore (a high school student!!). Both proofs use the Chevalley-Warning theorem, but in quite intricate and ingenious ways.

For any positive integer k, define $G(n,d) = (Z_n)^d$, the product of d copies of the cyclic group of order n, and consider lengths of sequences without an n-term zero sum subsequence: let us put $f(n,d)$ for the maximal length of such a sequence. Analogues of the above sequences with $\{0,1\}$-coordinates give

$$f(n,d) \geq 2^d(n-1).$$

In 1973 Heiko Harborth established the (much larger) upper bound

$$f(n,d) \leq n^d(n-1).$$

Harborth also computed $G(3,3) = 18 > 2^3(3-1)$: i.e., in this case the "obvious" examples do not have maximal length! It seems that the computation of $G(n,3)$ for all n – or still more, of $G(n,d)$ for all d – would be a significant achievement.

2. The Combinatorial Nullstellensatz

In this section we describe a celebrated result of Noga Alon that has served as a powerful technical tool and organizing principle in combinatorics and additive number theory. Recall that in §2 we considered the evaluation map

$$\Phi : k[t_1, \ldots, t_n] \to \mathrm{Map}(k^n, k)$$

and showed that when k is a finite field of order q, $\mathrm{Ker}\,\Phi$ is the ideal I_0 generated by $t_1^q - t_1, \ldots, t_n^q - t_n$. We showed this in a somewhat unorthodox way: first, by an explicit construction we saw that Φ is surjective. Further, clearly $\mathrm{Ker}\,\Phi$ contains

I_0, and thus by a counting argument $\operatorname{Ker} \Phi = I_0$. At the time we remarked that a more standard approach is to show that no nonzero reduced polynomial evaluates to the zero function by an induction on the number of variables. We are now in a position to want that argument, and in fact the following (mild) generalization.

LEMMA 15.5. *(Alon-Tarsi [**AT92**]) Let k be a field, $n \in \mathbb{Z}^+$, and $f(t) \in k[t] = k[t_1, \ldots, t_n]$; for $1 \leq i \leq n$, let d_i be the t_i-degree of f, let S_i be a subset of k with $\#S_i > d_i$, and let $S = \prod_{i=1}^n S_i$. If $f(x) = 0$ for all $x \in S$, then $f = 0$.*

PROOF. We go by induction on n. The case $n = 1$ is truly basic: a nonzero univariate polynomial over a field has no more roots than its degree. Now suppose $n \geq 2$ and that the result holds for polynomials in $n-1$ variables. The basic idea is the identity $k[t_1, \ldots, t_{n-1}, t_n] = k[t_1, \ldots, t_{n-1}][t_n]$: thus we write

$$f = \sum_{i=0}^{d_n} f_i(t_1, \ldots, t_{n-1}) t_n^i$$

with $f_i \in k[t_1, \ldots, t_{n-1}]$. If $(x_1, \ldots, x_{n-1}) \in k^{n-1}$, the polynomial $f(x_1, \ldots, x_{n-1}, t_n) \in k[t_n]$ has degree at most d_n and vanishes for all $\#S_n > d_n$ elements $x_n \in S_n$, so it is identically zero, i.e., $f_i(x_1, \ldots, x_{n-1}) = 0$ for all $0 \leq i \leq t_n$. By induction, each $f_i(t_1, \ldots, t_{n-1})$ is the zero polynomial and thus f is the zero polynomial. □

And now the main attraction.

THEOREM 15.6. *(Combinatorial Nullstellensatz) Let k be a field, $S_1, \ldots, S_n$ be nonempty finite subsets of k, and put $S = \prod_{i=1}^n S_i$. For $1 \leq i \leq n$, put*

$$g_i(t_i) = \prod_{s_i \in S_i} (t_i - s_i) \in k[t] = k[t_1, \ldots, t_n].$$

Suppose that for all $s = (s_1, \ldots, s_n) \in S$ we have $f(s) = 0$. Then there are polynomials $h_1(t_1, \ldots, t_n), \ldots, h_n(t_1, \ldots, t_n)$ such that

$$f = \sum_{i=1}^n h_i g_i.$$

PROOF. For $1 \leq i \leq n$, put $d_i = \#S_i - 1$; we may write

$$g_i(t_i) = t_i^{d_i+1} - \sum_{j=0}^{d_i} g_{ij} t_i^j.$$

Observe that if $s_i \in S_i$, then $g_i(x_i) = 0$, i.e.,

$$x_i^{d_i+1} = \sum_{j=0}^{d_i} g_{ij} x_i^j. \tag{48}$$

Let $\overline{f}$ be the polynomial obtained from f by writing f as a sum of monomials and repeatedly substituting each instance of $t_i^{e_i}$ with $e_i > d_i$ with a k-linear combination of smaller powers of t_i using (48). Then the **reduced** polynomial $\overline{f}$ has degree at most d_i in t_i and $f - \overline{f}$ is of the form $\sum_{i=1}^n h_i g_i$. Further, for all $s = (s_1, \ldots, s_n) \in S$, $\overline{f}(s) = f(s) = 0$. Thus Lemma 15.5 applies to give $\overline{f} = 0$ and thus $f = \sum_{i=1}^n h_i g_i$. □

We hope the reader has noticed that the proof of Theorem 15.6 bears more than a passing resemblance to our first proof of Warning's Theorem.

Example: Let $k = \mathbb{F}_q$ be a finite field, and take $S_1 = \ldots = S_n = \mathbb{F}_q$. Then for $1 \leq i \leq n$, $g_i = t_i^q - t_i$. Applying the Combinatorial Nullstellensatz, we see that a polynomial f which vanishes at every point in $\mathbb{F}_q^n$ lies in the ideal $I_0 = \langle t_1^q - t_1, \ldots, t_n^q - t_n \rangle$. That is, we have yet again computed the kernel of the evaluation map Φ.

Exercise: a) Show that, in the notation of the proof of Theorem 15.6, the polynomials $h_1, \ldots, h_n$ satisfy $\deg h_i \leq \deg f - \deg g_i$ for all $1 \leq i \leq n$.
b) Show that the coefficients of $h_1, \ldots, h_n$ lie in the subring of k generated by the coefficients of $f, g_1, \ldots, g_n$.

COROLLARY 15.7. *(Polynomial Method) Let k be a field, $n \in \mathbb{Z}^+$, $a_1, \ldots, a_n \in \mathbb{N}$, and let $f \in k[t] = k[t_1, \ldots, t_n]$. We suppose:*
(i) $\deg f = a_1 + \ldots + a_n$.
(ii) The coefficient of $t_1^{a_1} \cdots t_n^{a_n}$ in f is nonzero.
Then, for any subsets $S_1, \ldots, S_n$ of k with $\#S_i > a_i$ for $1 \leq i \leq n$, there is $s = (s_1, \ldots, s_n) \in S = \prod_{i=1}^n S_i$ such that $f(s) \neq 0$.

PROOF. It is no loss of generality to assume that $\#S_i = a_i + 1$ for all i, and we do so. We will show that if (i) holds and $f|_S \equiv 0$, then (ii) does *not* hold, i.e., the coefficient of $t_1^{a_1} \cdots t_n^{a_n}$ in f is 0.

Define, for all $1 \leq i \leq n$, $g_i(t_i) = \prod_{s_i \in S_i} t_i - s_i$. By Theorem 15.6 and the preceding exercise, there are $h_1, \ldots, h_n \in k[t]$ such that

$$f = \sum_{i=1}^n h_i g_i$$

and

$$\deg h_i \leq (a_1 + \ldots + a_n) - \deg g_i, \ \forall 1 \leq i \leq n,$$

so

$$\deg h_i g_i \leq \deg f. \tag{49}$$

Thus if $h_i g_i$ contains any monomial of degree $\deg f$, such a monomial would be of maximal degree in $h_i g_i = h_i \prod_{s_i \in S_i} (t_i - s_i)$ and thus be divisible by $t_i^{a_i+1}$. It follows that for all i, the coefficient of $t_1^{a_1} \cdots t_n^{a_n}$ in $h_i g_i$ is zero, hence the coefficient of $t_1^{a_1} \cdots t_n^{a_n}$ in f is zero. □

3. The Cauchy-Davenport Theorem

For nonempty subsets A, B of a group $(G, +)$,[2] we define the **sumset**

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

Exercise: Establish the **trivial bound**

$$\#(A + B) \geq \max \#A, \#B. \tag{50}$$

[2]In additive combinatorics, it is standard to write even non-commutative groups additively.

THEOREM 15.8. *(Cauchy-Davenport) Let p be a prime number. For nonempty subsets A, B of $\mathbb{Z}/p\mathbb{Z}$, we have*

$$\#(A+B) \geq \min(p, \#A + \#B - 1).$$

PROOF. Case 0: We may assume $\#A + \#B > 2$.
Case 1: Suppose $\#A+\#B > p$. Then for all $x \in \mathbb{Z}/p\mathbb{Z}$ we must have $A \cap (x-B) \neq \varnothing$, hence $x \in A + B$. Thus $A + B = \mathbb{Z}/p\mathbb{Z}$ and $\#(A + B) = p$.
Case 2: Suppose $\#A+\#B \leq p$ and, seeking a contradiction, that $\#(A+B) \leq \#A+\#B-2$. Let $C \subset \mathbb{Z}/p\mathbb{Z}$ be such that $A+B \subset C \subset \mathbb{Z}/p\mathbb{Z}$ and $\#C = \#A+\#B-2$. The polynomial

$$f(t_1, t_2) = \prod_{c \in C} (t_1 + t_2 - c) \in \mathbb{F}_p[t_1, t_2]$$

vanishes identically on $A \times B$. Let $d_1 = \#A - 1$, $d_2 = \#B - 1$; then the coefficient of $t_1^{d_1} t_2^{d_2}$ in f is the binomial coefficient $\binom{\#A+\#B-2}{\#A-1}$, which is nonzero in $\mathbb{Z}/p\mathbb{Z}$ since $\#A + \#B - 2 < p$. This contradicts Corollary **??**. □

Exercise: Show that the Cauchy-Davenport Theorem is *sharp*, in the following sense: for any prime p and integers a, b with $1 \leq a, b, \leq p$, there are subsets $A, B \subset \mathbb{Z}/p\mathbb{Z}$ with $\#A = a$, $\#B = b$ and $\#A + B = \min(p, \#A + \#B - 1)$.

G. Károlyi and (slightly later, but independently) J.P. Wheeler gave the following interesting generalization: for a (not necessarily abelian) group $(G, +)$, we define $p(G)$ to be the least order of a nonzero element of G, or ∞ if G has no nonzero elements of finite order.

Exercise: a) Show that $p(\mathbb{Z}/p\mathbb{Z}) = p$.
b) If G is finite, show that $p(G)$ is the least prime divisor of $\#G$.
c) Show that in any group G, if $p(G) < \infty$, then $p(G)$ is a prime number.

THEOREM 15.9. *(Károlyi-Wheeler* [**Ká05**], [**Wh12**]*) For nonempty subsets A, B of a finite group G,*

$$\#(A+B) \geq \min(p(G), \#A + \#B - 1).$$

Here is a very rough sketch of Wheeler's proof of Theorem 15.9: if $\#G$ is even then $p(G) = 2$ and the trivial bound (50) is sufficient. If $\#G$ is odd, then by the **Feit-Thompson Theorem** G is solvable (!!). Thus there is a normal subgroup H of G of prime index, whence an isomorphism $G/H \cong \mathbb{Z}/p\mathbb{Z}$. By an inductive argument (it takes about a page), one reduces to the Cauchy-Davenport Theorem.

CHAPTER 16

Dirichlet Series

1. Introduction

In considering the arithmetical functions $f : \mathbb{N} \to \mathbb{C}$ as a ring under pointwise addition and "convolution":

$$f * g(n) = \sum_{d_1 d_2 = n} f(d_1) g(d_2),$$

we employed that old dirty trick of abstract algebra. Namely, we introduced an algebraic structure without any motivation and patiently explored its consequences until we got to a result that we found useful (Möbius Inversion), which gave a sort of retroactive motivation for the definition of convolution.

This definition could have been given to an 18th or early 19th century mathematical audience, but it would not have been very popular: probably they would not have been comfortable with the Humpty Dumpty-esque redefinition of multiplication.[1] Mathematics at that time did have commutative rings: rings of numbers, of matrices, of functions, but not rings with a "funny" multiplication operation defined for no better reason than mathematical pragmatism.

So despite the fact that we have shown that the convolution product is a useful operation on arithmetical functions, one can still ask what $f * g$ "really is." There are (at least) two possible kinds of answers to this question: one would be to create a general theory of convolution products of which this product is an example and there are other familiar examples. Another would be to show how $f * g$ is somehow a more familiar multiplication operation, albeit in disguise.

To try to take the first approach, consider a more general setup: let $(M, \bullet)$ be a commutative monoid. Recall from the first homework assignment that this means that M is a set endowed with a binary operation $\bullet$ which is associative, commutative, and has an identity element, say e: $e \bullet m = m \bullet e = m$ for all $m \in M$. Now consider the set of all functions $f : M \to \mathbb{C}$. We can add functions in the obvious "pointwise" way:

$$(f + g)(m) := f(m) + g(m).$$

We could also multiply them pointwise, but we choose to do something else, defining

$$(f * g)(m) := \sum_{d_1 \bullet d_2 = m} f(d_1) g(d_2).$$

But not so fast! For this definition to make sense, we either need some assurance that for all $m \in M$ the set of all pairs d_1, d_2 such that $d_1 \cdot d_2 = m$ is finite (so

[1] Recall that Lewis Carroll – or rather Charles L. Dodgson (1832-1898) – was a mathematician.

the sum is a finite sum), or else some analytical means of making sense of the sum when it is infinite. But let us just give three examples:

Example 1: $(M, \bullet) = (\mathbb{Z}^+, \cdot)$. This is the example we started with – and of course the set of pairs of positive integers whose product is a given positive integer is finite.

Example 2: $(M, \bullet) = (\mathbb{N}, +)$. This is the "additive" version of the previous example:

$$(f * g)(n) = \sum_{i+j=n} f(i)g(j).$$

Of course this sum is finite: indeed, for $n \in \mathbb{N}$ it has exactly $n+1$ terms. As we shall see shortly, this "additive convolution" is closely related to the Cauchy product of infinite series.

Example 3: $(M, \bullet) = (\mathbb{R}, +)$. Here we have seem to have a problem, because for functions $f, g : \mathbb{R} \to \mathbb{C}$, we are defining

$$(f * g)(x) = \sum_{d_1+d_2=x} f(d_1)g(d_2) = \sum_{y \in \mathbb{R}} f(x-y)g(y),$$

and although it is possible to define a sum over all real numbers, it turns out never to converge unless f and g are zero for the vast majority of their values.[2] However, there is a well-known replacement for a "sum over all real numbers": the integral. So one should probably define

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy.$$

Here still one needs some conditions on f and g to ensure convergence of this "improper" integral. It is a basic result of analysis that if

$$\int_{-\infty}^{\infty} |f| < \infty, \ \int_{-\infty}^{\infty} |g| < \infty,$$

then the convolution product is well-defined. The convolution is an all-important operation in harmonic analysis: roughly speaking, it provides a way of "mixing together" two functions. Like any averaging process, it often happens that $f * g$ has nicer properties than its component functions: for instance, when f and g are absolutely integrable in the above sense, then $f * g$ is not only absolutely integrable but also continuous.

The most important property of this convolution is its behavior with respect to the **Fourier transform**: for a function $f : \mathbb{R} \to \mathbb{C}$ and $y \in \mathbb{R}$, one defines

$$\hat{f}(x) = \int_{-\infty}^{\infty} f(y)e^{-2\pi ix}dy.$$

Then one has the following identity:

$$\widehat{f * g} = \hat{f} \cdot \hat{g}.$$

[2]More precisely, if S is an arbitrary set of real numbers, it makes sense to define $\sum_{x_i \in S} = x$ if for all $\epsilon > 0$, there exists a finite subset $T \subset S$ such that for all finite subsets $T' \supset T$ we have $|\sum_{x_i \in T} x_i - x| < \epsilon$. (This is a special case of a **Moore-Smith limit**.) It can be shown that such a sum can only converge if the set of indices i such that $x_i \neq 0$ is finite or countably infinite.

In other words, there is a natural type of "transform" $f \mapsto \hat{f}$ under which the convolution becomes the more usual pointwise product.

Now the question becomes: is there some similar type of "transform" $f \mapsto \hat{f}$ which carries functions $f : M \to \mathbb{C}$ to some other space of functions and under which the convolution product becomes the pointwise product?

The answer is well-known to be "yes" if M is a locally compact abelian group (e.g. $\mathbb{Z}, \mathbb{Z}/N\mathbb{Z}, \mathbb{R}^n, \ldots$), and the construction is in fact rather similar to the above: this is the setting of **abstract Fourier analysis**. But our Examples 1 and 2 involve monoids that are not groups, so what we are looking for is not *exactly* a Fourier transform. So let us come back to earth by looking again at Examples 1 and 2.

In the case of Example 2, the construction we are looking for is just:

$$f \iff \{f(n)\}_{n=0}^{\infty} \mapsto F(x) = \sum_{n=0}^{\infty} f(n)x^n.$$

That is, to the sequence $\{f(n)\}$ we associate the corresponding **power series** $F(x) = \sum_n f(n)x^n$. One can look at this construction both *formally* and *analytically*.

The formal construction is purely algebraic: the ring of formal power series $\mathbb{C}[[t]]$ consists of all expressions of the form $\sum_{n=0}^{\infty} a_n x^n$ where the a_n's are complex numbers. We define addition and multiplication in the "obvious ways":

$$\sum_{n=0}^{\infty} a_n x^n + \sum_{n=0}^{\infty} b_n x^n := \sum_{n=0}^{\infty} (a_n + b_n) x^n,$$

$$(\sum_{n=0}^{\infty} a_n x^n)(\sum_{n=0}^{\infty} b_n x^n) := \sum_{n=0}^{\infty} (a_0 b_n + a_1 b_{n-1} + \ldots + a_n b_0) x^n.$$

The latter definition seems obvious because it is consistent with the way we multiply polynomials, and indeed the polynomials $\mathbb{C}[t]$ sit inside $\mathbb{C}[[t]]$ as the subring of all formal expressions $\sum_n a_n x^n$ with $a_n = 0$ for all sufficiently large n. Now note that this definition of multiplication is just the convolution product in the additive monoid $(\mathbb{N}, +)$:

$$a_0 b_n + \ldots + a_n b_0 = (a * b)(n).$$

It is not immediately clear that anything has been gained. For instance, it is, technically, not for free that this multiplication law of formal power series is associative (although of course this is easy to check). Nevertheless, one should not underestimate the value of this purely formal approach. Famously, there are many nontrivial results about sequences f_n which can be proved just by simple algebraic manipulations of the "generating function" $F(x) = \sum_n f_n x^n$. For example:

THEOREM 16.1. *Let $a_1, \ldots, a_k$ be a coprime set of positive integers, and define $r(N)$ to be the number of solutions to the equation*

$$x_1 a_1 + \ldots + x_k a_k = N$$

in non-negative integers $x_1, \dots, x_k$. Then as $N \to \infty$,

$$r(N) \sim \frac{N^{k-1}}{(k-1)!(a_1 \cdots a_k)}.$$

Nevertheless we also have and need an *analytic* theory of power series, i.e., of the study of properties of $F(x) = \sum_n a_n x^n$ viewed as a function of the complex variable x. This theory famously works out very nicely, and can be summarized as follows:

THEOREM 16.2. *(Theory of power series) Let $\sum_n a_n x^n$ be a power series with complex coefficients. Let $R = (\limsup_n |a_n|^{\frac{1}{n}})^{-1}$. Then:*
a) The series converges absolutely for all $x \in \mathbb{C}$ with $|x| < R$, and diverges – indeed, the general term tends to infinity in modulus – for all x with $|x| > R$.
b) The convergence is uniform on compact subsets of the open disk of radius R (about 0), from which it follows that $F(x)$ is a complex analytic function on this disk.
c) If two power series $F(x) = \sum_n a_n x^n$, $G(x) = \sum_n b_n x^n$ are defined and equal for all x in some open disk of radius $R > 0$, then $a_n = b_n$ for all n.

In particular, it follows from Cauchy's theory of products of absolutely convergent series that if $F(x) = \sum_n a_n x^n$ and $G(x) = \sum_n b_n x^n$ are two power series convergent on some disk of radius $R > 0$, then on this disk the function FG – the product of F and G in the usual sense – is given by the power series $\sum_n (a * b)(n) x^n$. In other words, with suitable growth conditions on the sequences, we get that the product of the transforms is the transform of the convolutions, as advertised.

Now we return to the case of interest: $(M, \bullet) = (\mathbb{Z}^+, \cdot)$. The transform that does the trick is $f \mapsto D(f, s)$, where $D(f, s)$ is the **formal Dirichlet series**

$$D(f, s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s}.$$

To justify this, suppose we try to formally multiply out

$$D(f, s) D(g, s) = \left(\sum_{m=1}^{\infty} \frac{f(m)}{m^s} \right) \left(\sum_{n=1}^{\infty} \frac{g(n)}{n^s} \right).$$

We will get one term for each pair (m, n) of non-negative integers, so the product is (at least formally) equal to

$$\sum_{(m,n)} \frac{f(m)g(n)}{m^s n^s} = \sum_{(m,n)} \frac{f(m)g(n)}{(mn)^s},$$

where in both sums m and n range over over all positive integers. To make a Dirichlet series out of this, we need to collect all the terms with a given denominator, say N^s. The only way to get 1 in the denominator is to have $m = n = 1$, so the first term is $\frac{f(1)g(1)}{1^s}$. Now to get a 2 in the denominator we could have $m = 1,\ n = 2$ – giving the term $\frac{f(1)g(2)}{2^s}$ – or also $m = 2, n = 1$ – giving the term $\frac{f(2)g(1)}{2^s}$, so all in all the numerator of the "2^s-term" is $f(1)g(2) + f(2)g(1)$.

Aha. In general, to collect all the terms with a given denominator N^s in the

product involves summing over all expressions $f(m)g(n)$ with $mn = N$. In other words, we have the following formal identity:

$$D(f,s) \cdot D(g,s) = (\sum_{n=1}^{\infty} \frac{f(n)}{n^s})(\sum_{n=1}^{\infty} \frac{g(n)}{n^s}) = \sum_{n=1}^{\infty} \frac{\sum_{d|n} f(d)g(n/d)}{n^s} = D(f*g,s).$$

Thus we have attained our goal: under the "transformation" which associates to an arithmetical function its Dirichlet series $D(f,s)$, Dirichlet convolution of arithmetical functions becomes the usual multiplication of functions!

There are now several stages in the theory of Dirichlet series:

Step 1: Explore the purely formal consequences: that is, that identities involving convolution and inversion of arithmetical functions come out much more cleanly on the Dirichlet series side.

Step 2: Develop the theory of $D(f,s)$ as a function of a complex variable s. It is rather easy to tell when the series $D(f,s)$ is *absolutely* convergent. In particular, with suitable growth conditions on $f(n)$ and $g(n)$, we can see that

$$D(f,s)D(g,s) = D(f*g,s)$$

holds not just formally but also as an equality of functions of a complex variable. In particular, this leads to an "analytic proof" of the Möbius Inversion Formula.

On the other hand, unlike power series there can be a region of the complex plane with nonempty interior in which the Dirichlet series $D(f,s)$ is only conditionally convergent (that is, convergent but not absolutely convergent). We will present, without proofs, the basic results on this more delicate convergence theory.

In basic analysis we learn to abjure conditionally convergent series, but they lie at the heart of analytic number theory. In particular, in order to prove Dirichlet's theorem on arithmetic progressions one studies the Dirichlet series $L(\chi,s)$ attached to a **Dirichlet character** χ (a special kind of arithmetical function we will define later on), and it is extremely important that for all $\chi \neq \mathbf{1}$, there is a "critical strip" in the complex plane for which $L(\chi,s)$ is only conditionally convergent. We will derive this using the assumed results about conditional convergence of Dirichlet series and a convergence test, **Dirichlet's test**, from advanced calculus.[3] Finally, as an example of how much more content and subtlety lies in conditionally convergent series, we will use Dirichlet series to give an analytic continuation of the zeta function to the right half-plane (complex numbers with positive real part), which allows for a rigorous and concrete statement of the Riemann hypothesis.

2. Some Dirichlet Series Identities

Example 1: If $f = \mathbf{1}$ is the constant function 1, then by definition $D(\mathbf{1},s)$ is what is probably the most single important function in all of mathematics, the **Riemann zeta function**:

$$\zeta(s) = D(\mathbf{1},s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

[3]P.G.L. Dirichlet propounded his convergence test with this application in mind.

Example 2: Let $f(n) = d(n)$, the divisor function, so $D(d,s) = \sum_n \frac{d(n)}{n^s}$. But we also know that $d = \mathbf{1} * \mathbf{1}$. On Dirichlet series this means that we multiply: so that $D(d,s) = D(\mathbf{1},s)D(\mathbf{1},s)$, and we get that

$$D(d,s) = \zeta(s) \cdot \zeta(s) = \zeta^2(s).$$

Example 3: Since $\delta(1) = 1$ and $\delta(n) = 0$ for all $n > 1$, we have $D(\delta,s) = \frac{1}{1^s} + \sum_{n=2}^{\infty} \frac{0}{n^s} = 1$. Thus the Dirichlet series of the δ – the multiplicative identity for convolution – is just the constant function 1, the multiplicative identity in the "usual" sense of multiplication functions.

Example 4: What is $D(\mu,s)$? Since $\mu * \mathbf{1} = \delta$, we must have

$$1 = D(\iota,s) = D(\mu,s)D(\mathbf{1},s) = D(\mu,s)\zeta(s),$$

so

$$D(\mu,s) = \frac{1}{\zeta(s)}.$$

Probably this is the most important such identity: it relates *combinatorial* methods (the Möbius function is closely related to the inclusion-exclusion principle) to *analytical* methods. More on this later.

We record without proof the following further identities, whose derivations are similarly straightforward. Some notational reminders: we write ι for the function $n \mapsto n$; ι_k for the function $n \mapsto n^k$; and λ for the function $n \mapsto (-1)^{\Omega(n)}$, where $\Omega(n)$ is the number of prime divisors of n counted with multiplicity.

$$D(\iota,s) = \zeta(s-1).$$

$$D(\iota_k,s) = \zeta(s-k).$$

$$D(\sigma,s) = \zeta(s)\zeta(s-1).$$

$$D(\sigma_k,s) = \zeta(s)\zeta(s-k).$$

$$D(\varphi,s) = \frac{\zeta(s-1)}{\zeta(s)}.$$

$$D(\lambda,s) = \frac{\zeta(2s)}{\zeta(s)}.$$

3. Euler Products

Our first task is to make formal sense of an infinite product of infinite series, which is unfortunately somewhat technical. Suppose that we have an infinite indexing set P and for each element of p of P an infinite series whose first term is 1:

$$\sum_{n=0}^{\infty} a_{p,n} = 1 + a_{p,1} + a_{p,2} + \dots.$$

Then by the infinite product $\prod_{p \in P} \sum_n a_{p,n}$ we mean an infinite series whose terms are indexed by the infinite direct sum $\mathcal{T} = \bigoplus_{p \in P} \mathbb{N}$. Otherwise put, an element

$t \in \mathcal{T}$ is just a function $t : P \to \mathbb{N}$ such that $t(p) = 0$ for all but finitely many p in P.[4] Then by $\prod_{p \in P} \sum_n a_{p,n}$ we mean the formal infinite series

$$\sum_{t \in \mathcal{T}} \prod_{p \in P} a_{p,t(p)}.$$

Note well that for each t, since $t(p) = 0$ except for finitely many p and since $a_{p,0} = 1$ for all p, the product $\prod_{p \in P} a_{p,t(p)}$ is really a finite product. Thus the series is well-defined "formally" – that is, merely in order to write it down, no notion of limit of an infinite process is involved.

Let us informally summarize the preceding: to make sense of a formal infinite product of the form

$$\prod_p \left(1 + a_{p,1} + a_{p,2} + \ldots + a_{p,n} + \ldots\right),$$

we give ourselves one term for each possible product of one term from the first series, one term from the second series, and so forth, but we are only allowed to choose a term which is different from the $a_{p,0} = 1$ term finitely many times.

With that out of the way, recall that when developing the theory of arithmetical functions, we found ourselves in much better shape under the hypothesis of **multiplicativity**. It is natural to ask what purchase we gain on $D(f, s)$ by assuming the multiplicativity of f. The answer is that multiplicativity of f is equivalent to the following formal identity:

$$D(f, s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s} = \prod_p \left(1 + \frac{f(p)}{p^s} + \frac{f(p^2)}{p^{2s}} + \ldots\right). \tag{51}$$

Here the product extends over all primes. The fact that this identity holds (as an identity of formal series) follows from the uniqueness of the prime power factorization of positive integers.

An expression as in (51) is called an **Euler product expansion**. If f is moreover completely multiplicative, then $\frac{f(p^k)}{p^{ks}} = (\frac{f(p)}{p^s})^k$, and each factor in the product is a geometric series with ratio $\frac{f(p)}{p^s}$, so we get

$$D(f, s) = \prod_p \left(1 - \frac{f(p)}{p^s}\right)^{-1}.$$

In particular $f = \mathbf{1}$ is certainly completely multiplicative, so we get the identity

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1},$$

which we used in our study of the primes. We also get

$$\sum_{n=1}^{\infty} \frac{\mu(n)}{n^s} = \frac{1}{\zeta(s)} = \prod_p \left(1 - \frac{1}{p^s}\right), \tag{52}$$

[4]The property that $t(p) = 0$ except on a finite set is, by definition, what distinguishes the infinite direct sum from the infinite direct product.

and, plugging in $s = 2$,

$$\frac{6}{\pi^2} = \frac{1}{\zeta(2)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^2} = \prod_p \left(1 - \frac{1}{p^2}\right).$$

But not so fast! We changed the game here: so far (52) expresses a *formal* identity of Dirichlet series. In order to be able to plug in a value of s, we need to discuss the *convergence* properties of Dirichlet series and Euler products. In particular, since we did not put any particular ordering on our formal infinite product, in order for the sum to be meaningful we need the series involved to be *absolutely* convergent. It is therefore to this topic that we now turn.

4. Absolute Convergence of Dirichlet Series

Let us first study the *absolute convergence* of Dirichlet series $\sum_n \frac{a_n}{n^s}$. That is, we will look instead at the series $\sum_n \frac{|a_n|}{n^\sigma}$, where $s = \sigma + it$.[5]

THEOREM 16.3. *Suppose a Dirichlet series $D(s) = \sum_n \frac{a_n}{n^s}$ is absolutely convergent at some complex number $s_0 = \sigma_0 + it_0$. Then it is also absolutely convergent at all complex numbers s with $\sigma = \Re(s) > s_0$.*

PROOF. If $\sigma = \Re(s) > \sigma_0 = \Re(s_0)$, then $n^{-\sigma} > n^{\sigma_0}$ for all $n \in \mathbb{Z}^+$, so

$$\sum_{n=1}^{\infty} |\frac{a_n}{n^s}| = \sum_{n=1}^{\infty} \frac{|a_n|}{n^\sigma} \leq \sum_{n=1}^{\infty} \frac{|a_n|}{n^{\sigma_0}} = \sum_{n=1}^{\infty} |\frac{a_n}{n^{s_0}}| < \infty.$$

□

It follows that the **domain of absolute convergence** of a Dirichlet series $D(f, s)$ is one of the following:

(i) The empty set. (I.e., for no s does the series absolutely converge.)
(ii) All of $\mathbb{C}$.
(iii) An open half-plane of the form $\Re s > S$.
(iv) A closed half-plane of the form $\Re s \geq S$.

Notice that in all cases, there is a unique $\sigma_{\text{ac}} \in [-\infty, \infty]$ such that:

(AAC1) For all s with $\Re(s) > \sigma_{\text{ac}}$, $D(s)$ is absolutely convergent.
(AAC2) For all s with $\Re(s) < \sigma_{\text{ac}}$, $D(s)$ is **not** absolutely convergent.

This unique σ_{ac} is called the **abscissa of absolute convergence** of $D(s)$.

Example 1 (Type i): $D(s) = \sum_n \frac{2^n}{n^s}$.
This series does not converge (absolutely or otherwise) for any $s \in \mathbb{C}$: no matter what s is, $|2^n \cdot n^{-s}| \to \infty$: exponentials grow faster than power functions. So $\sigma_{\text{ac}} = \infty$.

Example 2 (Type ii): A trivial example is the zero series – $a_n = 0$ for all n, or

[5]In other words, for a complex number s we write σ for its real part and t for its imaginary part. This seemingly unlikely notation was introduced in a fundamental paper of Riemann, and remains standard to this day.

for that matter, any series with $a_n = 0$ for all sufficiently large n: these give finite sums. Or we could take $a_n = 2^{-n}$ and now the series converges absolutely independent of s. So $\sigma_{\text{ac}} = -\infty$.

Example 3 (Type iii): $\zeta(s) = \frac{1}{n^s}$ is absolutely convergent for $s \in (1, \infty)$. So $\sigma_{\text{ac}} = 1$.

Example 4 (Type iv): For $a_n = \frac{1}{(\log n)^2}$, the domain of absolute convergence is $[1, \infty)$.

The following result gives a sufficient condition for $\sigma_{\text{ac}} = 1$:

PROPOSITION 16.4. *Let $D(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$ be a Dirichlet series.*
a) Suppose that there is $M \in \mathbb{R}$ such that $|a_n| \leq M$ for all n. Then $\sigma_{\text{ac}} \leq 1$.
*b) Suppose the sequence a_n does **not** tend to 0. Then $\sigma_{\text{ac}} \geq 1$.*
c) In particular if the sequence a_n is bounded but not convergent to 0, then $\sigma_{\text{ac}} = 1$.

PROOF. a) Suppose $|a_n| \leq M$ for all n and also that $\sigma = \Re(s) > 1$. Then

$$\sum_n |\frac{a_n}{n^s}| \leq M \sum_n \frac{1}{n^\sigma} = M\zeta(\sigma) < \infty.$$

b) The Dirichlet series at 0 is $\sum_n \frac{a_n}{n^0} = \sum_n a_n$. Of course this series can only be convergent (absolutely or otherwise) if $a_n \to 0$. Part c) follows immediately from a) and b). □

Definition: We say that an arithmetic function $a_n : \mathbb{Z}^+ \to \mathbb{C}$ has **polynomial growth of order N** if there exist positive real numbers C and N such that $|a_n| \leq Cn^N$ for all $n \in \mathbb{Z}^+$. We say that a function has **polynomial growth** if it has polynomial growth of order N for some $N \in \mathbb{R}^+$.

PROPOSITION 16.5. *Suppose $\{a_n\}$ has polynomial growth of order N. Then the associated Dirichlet series $D(s) = \frac{a_n}{n^s}$ has $\sigma_{\text{ac}} \leq N + 1$.*

PROOF. By hypothesis, there exists C such that $|a_n| \leq Cn^N$ for all $n \in \mathbb{Z}^+$. If $\sigma = \Re(s) > N + 1$, then there exists $\epsilon > 0$ such that $\sigma > N + 1 + \epsilon$. Then

$$\sum_n |\frac{a_n}{n^\sigma}| \leq \sum_n \frac{|a_n|}{n^{N+1+\epsilon}} \leq C \sum_n \frac{n^N}{n^{N+1+\epsilon}} = C \sum_n \frac{1}{n^{1+\epsilon}} < \infty.$$

□

COROLLARY 16.6. *Let $f(n)$, $g(n)$ be arithmetical functions with polynomial growth of order N. Then*

$$D(f, s)D(g, s) = D(f * g, s)$$

is an equality of functions defined on $(N + 1, \infty)$.

This follows easily from the theory of absolute convergence and the Cauchy product.

THEOREM 16.7. *(Uniqueness Theorem) Let $f(n)$, $g(n)$ be arithmetical functions whose Dirichlet series are both absolutely convergent in the halfplane $\sigma = \Re(s) > \sigma_0$. Suppose there exists an infinite sequence s_k of complex numbers, with $\sigma_k = \Re(s_k) > \sigma_0$ for all k and $\sigma_k \to \infty$ such that $D(f, s_k) = D(g, s_k)$ for all k. Then $f(n) = g(n)$ for all n.*

PROOF. First we put $h(n) := f(n) - g(n)$, so that $D(h,s) = D(f,s) - D(g,s)$. Then our assumption is that $D(h, s_k) = 0$ for all k, and we wish to show that $h(n) = 0$ for all n.

So suppose not, and let N be the least n for which $h(n) \neq 0$. Then

$$D(h,s) = \sum_{n=N}^{\infty} \frac{h(n)}{n^s} = \frac{h(N)}{N^s} + \sum_{n=N+1}^{\infty} \frac{h(n)}{n^s},$$

so

$$h(N) = N^s D(h,s) - N^s \sum_{n=N+1}^{\infty} \frac{h(n)}{n^s}.$$

Taking now $s = s_k$ we have that for all $k \in \mathbb{Z}^+$,

$$h(N) = -N^{s_k} \sum_{n=N+1}^{\infty} \frac{h(n)}{n^{-s_k}}.$$

Fix a $\sigma > \sigma_0$, and choose a k such that $\sigma_k > \sigma$. Then

$$|h(N)| \leq N^{\sigma_k} \sum_{n=N+1}^{\infty} |h(n)| n^{-\sigma_k} \leq \frac{N^{\sigma_k}}{(N+1)^{\sigma_k - c}} \sum_{n=N+1}^{\infty} |h(n)| n^{-c} \leq C \left(\frac{N}{N+1} \right)^{\sigma_k},$$

for some constant C independent of n and k. Since N is a constant, letting $\sigma_k \to \infty$ the right hand side approaches 0, thus $h(N) = 0$, a contradiction. □

COROLLARY 16.8. *Let $D(s) = \sum_n \frac{a_n}{n^s}$ be a Dirichlet series with abscissca of absolute convergence σ_{ac}. Suppose that for some s with $\Re(s) > \sigma_{\mathrm{ac}}$ we have $D(s) = 0$. Then there exists a halfplane in which $D(s)$ is absolutely convergent and never zero.*

PROOF. If not, we have an infinite sequence $\{s_k\}$ of complex numbers, with real parts tending to infinity, such that $D(s_k) = 0$ for all k. By the Uniqueness Theorem this implies that $a_n = 0$ for all n and thus $D(s)$ is identically zero in its halfplane of absolute convergence, contrary to our assumption. □

COROLLARY 16.9. *(MIF for polynomially growing functions) If $f(n)$ is an arithmetical function with polynomial growth and $F(n) = \sum_{d|n} f(n)$, then $f(n) = \sum_{d|n} F(d)\mu(n/d)$.*

Surely this was the first known version of the Möbius inversion formulaa. As Hardy and Wright remark [**HW**], the "real" proof of MIF is the purely algebraic one we gave earlier, but viewing things in terms of "honest" functions has a certain appeal.

Moreover, the theory of absolute convergence of infinite products (see e.g. [**A1**, §11.5]) allows us to justify our formal Euler product expansions:

THEOREM 16.10. *(Theorem 11.7 of [**A1**]) Suppose that $D(f,s) = \sum_n \frac{f(n)}{n^s}$ converges absolutely for $\sigma > \sigma_{ac}$. If f is multiplicative we have an equality of functions*

$$D(f,s) = \prod_p \left(1 + \frac{f(p)}{p^s} + \frac{f(p^2)}{p^{2s}} + \dots \right),$$

valid for all s with $\Re(s) > \sigma_{ac}$. If f is completely multiplicative, this simplifies to

$$D(f,s) = \prod_p \left(1 - \frac{f(p)}{p^s}\right)^{-1}.$$

Euler products are ubiquitous in modern number theory: they play a prominent role in (e.g.!) the proof of Fermat's Last Theorem.

5. Conditional Convergence of Dirichlet Series

Let $D(f,s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$ be a Dirichlet series. We assume that the abscissa of absolute convergence σ_{ac} is finite.

THEOREM 16.11. *There exists a real number σ_c with the following properties:*
(i) If $\Re(s) > \sigma_c$, then $D(f,s)$ converges (not necessarily absolutely).
(ii) If $\Re(s) < \sigma_c$, then $D(f,s)$ diverges.

Because the proof of this result is already somewhat technical, we defer it until §$X.X$ on general Dirichlet series, where we will state and prove a yet stronger result.

Definition: σ_c is called the **abscissa of convergence**.

Contrary to the case of absolute convergence we make no claims about the convergence or divergence of $D(f,s)$ along the line $\Re(s) = \sigma$: this is quite complicated.

PROPOSITION 16.12. *We have*

$$0 \leq \sigma_{ac} - \sigma_c \leq 1.$$

PROOF. Since absolutely convergent series are convergent, we evidently must have $\sigma_{ac} \geq \sigma$. On the other hand, let $s = \sigma + it$ be a complex number such that $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$ converges. Of course this implies that $\frac{a_n}{n^s} \to 0$ as $n \to \infty$, and that in turn implies that there exists an N such that $n \geq N$ implies $|\frac{a_n}{n^s}| = \frac{|a_n|}{n^\sigma} \geq 1$. Now let s' be any complex number with real part $\sigma + 1 + \epsilon$ for any $\epsilon > 0$. Then for all $n \geq N$,

$$|\frac{a_n}{n^{s'}}| = \frac{|a_n|}{n^\sigma} \cdot \frac{1}{n^{1+\epsilon}} \leq \frac{1}{n^{1+\epsilon}},$$

so by comparison to a p-series with $p = 1 + \epsilon > 1$, $D(f,s')$ is absolutely convergent. □

It can be a delicate matter to show that a series is convergent but not absolutely convergent: there are comparatively few results that give criteria for this. The following one – sometimes encountered in an advanced calculus class – will serve us well.

PROPOSITION 16.13. *(Dirichlet's Test) Let $\{a_n\}$ be a sequnece of complex numbers and $\{b_n\}$ a sequence of real numbers. Suppose both of the following hold:*
(i) There exists a fixed M such that for all $N \in \mathbb{Z}^+$, $|\sum_{n=1}^{N} a_n| \leq M$ (bounded partial sums);
(ii) $b_1 \geq b_2 \geq \ldots \geq b_n \geq \ldots$ and $\lim_n b_n = 0$.
Then $\sum_{n=1}^{\infty} a_n b_n$ is convergent.

PROOF. Write S_N for $\sum_{n=1}^N$, so that by (i) we have $|S_N| \leq M$ for all N. Fix $\epsilon > 0$, and choose N such that $b_N < \frac{1}{\epsilon 2M}$. Then, for all $m, n \geq N$:

$$\begin{aligned}
|\sum_{k=m}^{n} a_k b_k| &= \sum_{k=m}^{n} (S_k - S_{k-1}) b_k| \\
&= |\sum_{k=m}^{n} S_k b_k - \sum_{k=m-1}^{n-1} S_k b_{k+1}| \\
&= |\sum_{k=m}^{n-1} S_k (b_k - b_{k+1}) + S_n b_n - S_{m-1} b_m| \\
&\leq \sum_{k=m}^{n-1} |S_k||b_k - b_{k+1}| + |S_n||b_n| + |S_{m-1}||b_m| \\
&\leq M \left(\sum_{k=m}^{n-1} |b_k - b_{k+1}| + |b_n| + |b_m| \right) = 2Mb_m \leq 2Mb_N < \epsilon.
\end{aligned}$$

Therefore the series satisfies the Cauchy criterion and hence converges.[6] □

THEOREM 16.14. *Let $\{a_n\}_{n=1}^{\infty}$ be a complex sequence.*
a) Suppose that the partial sums $\sum_{n=1}^N a_n$ are bounded. Then the Dirichlet series $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$ has $\sigma_c \leq 0$.
b) Assume in addition that a_n does not converge to 0. Then $\sigma_{ac} = 1$, $\sigma_c = 0$.

PROOF. By Proposition 16.4, $\sigma_{\mathrm{ac}} = 1$. For any real number $\sigma > 0$, by taking $b_n = \frac{1}{n^\sigma}$ the hypotheses of Proposition 16.13 are satisfied, so that $D(\sigma) = \sum_n \frac{a_n}{n^\sigma}$ converges. The smallest right open half-plane which contains all positive real numbers σ is of course $\Re(s) > 0$, so $\sigma \leq 0$. By Proposition 16.12 we have $1 = \sigma_{ac} \leq 1 + \sigma$, so we conclude that $\sigma = 0$. □

THEOREM 16.15. *(Theorem 11.11 of [**A1**]) A Dirichlet series $D(f, s)$ converges uniformly on compact subsets of the half-plane of convergence $\Re(s) > \sigma$.*

Suffice it to say that, in the theory of sequences of functions, "uniform convergence on compact subsets" is the magic incantation. As a consequence, we may differentiate and integrate Dirichlet series term-by-term. Also:

COROLLARY 16.16. *The function $D(f, s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s}$ defined by a Dirichlet series in its half-plane $\Re(s) > \sigma$ of convergence is complex analytic.*

6. Dirichlet Series with Nonnegative Coefficients

Suppose we are given a Dirichlet series $D(s) = \sum_n \frac{a_n}{n^s}$ with the property that for all n, a_n is real and non-negative. There is more to say about the analytic theory of such series. First, the non-negativity hypothesis ensures that for any real s, $D(s)$ is a series with non-negative terms, so its absolute convergence is equivalent to its convergence. Thus:

LEMMA 16.17. *For a Dirichlet series with non-negative real coefficients, the abscissae of convergence and absolute convergence coincide.*

[6] This type of argument is known as **summation by parts**.

Thus one of the major differences from the the theory of power series is eliminated for Dirichlet series with non-negative real coefficients. Another critical property of all complex power series is that the radius of convergence R is as large as conceivably possible, in that the function necessarily has a singularity somewhere on the boundary of the disk $|z - z_0| < R$ of convergence. This property need not be true for an arbitrary Dirichlet series. Indeed the series

$$D(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n+1)^s} = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \dots,$$

has $\sigma = 0$ but extends to an analytic function on the entire complex plane.[7] However:

THEOREM 16.18. *(Landau) Let $D(s) = \sum_n \frac{a_n}{n^s}$ be a Dirichlet series, with a_n real and non-negative for all n. Suppose that for a real number σ, $D(s)$ converges in the half-plane $\Re(s) > \sigma$, and that $D(s)$ extends to an analytic function in a neigborhood of σ. Then σ strictly exceeds the abscissa of convergence σ_c.*

Proof (Kedlaya): Suppose on the contrary that $D(s)$ extends to an analytic function on the disk $|s - \sigma| < \epsilon$, for some $\epsilon > 0$ but $\sigma = \sigma_c$. Choose $c \in (\sigma, \sigma + \epsilon/2)$, and write

$$D(s) = \sum_n a_n n^{-c} n^{c-s} = \sum_n a_n n^{-c} e^{(c-s)\log n}$$
$$= \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} \frac{a_n n^{-c} (\log n)^k}{k!} (c-s)^k.$$

Here we have a double series with all coefficients non-negative, so it must converge absolutely on the disk $|s-c| < \frac{\epsilon}{2}$. In particular, viewed as a Taylor series in $(c-s)$, this must be the Taylor series expansion of $D(s)$ at $s = c$. Since $D(s)$ is assumed to be holomorphic in the disk $|s - c| < \frac{\epsilon}{2}$, this Taylor series is convergent there. In particular, choosing any real number σ' with $\sigma - \frac{\epsilon}{2} < \sigma' < \sigma$, we have that $D(\sigma')$ is absolutely convergent. But this implies that the original Dirichlet series is convergent at σ', contradiction!

For example, it follows from Landau's theorem that the Riemann zeta function $\zeta(s) = \sum_n \frac{1}{n^s}$ must have a singularity at $s = 1$, since otherwise there would exist some $\sigma < 1$ such that the series converges in the entire half-plane $\Re(s) > \sigma$.

Of course this is a horrible illustration of the depth of Landau's theorem, since we used the fact that $\zeta(1) = \infty$ in order to compute the abscissa of convergence of the zeta function! We will see a much deeper application of Landau's theorem during the proof of Dirichlet's theorem on primes in arithmetic progressions.

7. Characters and L-Series

Let $f : \mathbb{Z}^+ \to \mathbb{C}$ be an arithmetic function.

Recall that f is said to be **completely mutliplicative** if $f(1) \neq 0$ and for all $a, b \in \mathbb{Z}$, $f(ab) = f(a)f(b)$. The conditions imply $f(1) = 1$.

[7] We will see a proof of the former statement shortly, but not the latter. More generally, it is true for the L-function associated to any primitive Dirichlet character.

For $N \in \mathbb{Z}^+$, we say a function f is N-**periodic** if it satisfies:

(P_N) For all $n \in \mathbb{Z}^+$, $f(n+N) = f(n)$.

An arithmetic function is **periodic** if it is N-periodic for some $N \in \mathbb{Z}^+$.

Remark: A function $f : \mathbb{Z} \to \mathbb{C}$ is said to be N-periodic if for all $n \in \mathbb{Z}$, $f(n+N) = f(n)$. It is easy to see that any N-periodic arithmetic function admits a unique extension to an N-periodic function with domain $\mathbb{Z}$.

Note that if f is N-periodic it is also kN-periodic for every $k \in \mathbb{Z}^+$. Conversely, we define the **period** P of a periodic function to be the least positive integer N such that f is N-periodic, then it is easy to see that f is N-periodic iff $P \mid N$.

Now we are ready to meet the object of our affections:

A **Dirichlet character** is a periodic completely multiplicative arithmetic function.[8]

Example: For an odd prime p, define $L_p : \mathbb{Z}^+ \to \mathbb{C}$ by $L_p(n) = (\frac{n}{p})$ (Legendre symbol). The period of L_p is p. Notice that $L_p(n) = \pm 1$ if n is prime to p, whereas $L_p(n) = 0$ if $\gcd(n,p) > 1$. This generalizes as follows:

THEOREM 16.19. *Let f be a Dirichlet character of period N.*
a) If $\gcd(n,N) = 1$, then $f(n)$ is a $\varphi(N)$th root of unity in $\mathbb{C}$ (hence $f(n) \neq 0$).
b) If $\gcd(n,N) > 1$, then $f(n) = 0$.

PROOF. Put $d = \gcd(n,N)$. Assume first that $\gcd(n,N) = 1$, so by Lagrange's Theorem $n^{\varphi(N)} \equiv 1 \pmod{N}$. Then:

$$f(n)^{\varphi(N)} = f(n^{\varphi(N)}) = f(1) = 1.$$

Next assume $d > 1$, and write $n = dn_1$, $N = dN_1$. By assumption N_1 properly divides N, so is strictly less than N. Then f is not N_1-periodic, so there exists $m \in \mathbb{Z}^+$ such that

$$f(m + N_1) - f(m) \neq 0.$$

On the other hand

$$f(d)\,(f(m+N_1) - f(m)) = f(dm+N) - f(dm) = f(dm) - f(dm) = 0,$$

so

$$f(n) = f(dn_1) = f(d)f(n_1) = 0 \cdot f(n_1) = 0.$$

□

7.1. Period N Dirichlet characters and characters on $U(N)$.

A fruitful perspective on the Legendre character $L(p)$ is that it is obtained from a certain homomorphism from the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ into the multiplicative group $\mathbb{C}^\times$ of complex numbers by extending to $L(0 \pmod{p}) := 0$. In fact all Dirichlet characters of a given period can be constructed in this way.

[8]Recall that by definition a multiplicative function is not identically zero, whence $f(1) = 1$.

We introduce some further notation: for $N \in \mathbb{Z}^+$, let $U(N) = (\mathbb{Z}/N\mathbb{Z})^\times$ be the unit group, a finite commutative group of order $\varphi(N)$. Let $X(N)$ be the group of characters of $U(N)$, i.e., the group homomorphisms $U(N) \to \mathbb{C}^\times$. We recall from [Algebra Handout 2.5, §4] that $X(N)$ is a finite commutative group whose order is is also $\varphi(N)$.[9]

PROPOSITION 16.20. *Let N be a positive integer. There is a bijective correspondence between Dirichlet characters with period N and elements of $X(N) = \mathrm{Hom}(U(N), \mathbb{C}^\times)$.*

PROOF. If $f : U(N) \to \mathbb{C}$ is a homomorphism, we extend it to a function from $f : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$ by defining $f(0) = 0$ on all residue classes which are not prime to N, and then define

$$\tilde{f}(n) := f(n \mod N).$$

In other words, if $q_N : \mathbb{Z} \to \mathbb{Z}/N\mathbb{Z}$ is the quotient map, then $\tilde{f} := f \circ q_N$.

Conversely, if $f : \mathbb{Z}^+ \to \mathbb{C}$ is a Dirichlet character mod N, then its extension to $\mathbb{Z}$ is N-periodic and therefore factors through $\overline{f} : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$.

It is easy to see that these two constructions are mutually inverse. □

For example, the function $\mathbf{1} : n \to 1$ for all n is the unique Dirichlet character of period 1. The character $\mathbf{1}$ is said to be **trivial**; all other Dirichlet characters are said to be **nontrivial**. Under the correspondence of Proposition 16.20 it corresponds to the unique homomorphism from the trivial group $\mathbb{Z}/1\mathbb{Z} \to \mathbb{C}$.

7.2. Examples.

Example (Principal character): For any $N \in \mathbb{Z}^+$, define $\xi_N : \mathbb{Z}^+ \to \mathbb{C}$ by

$$\xi_N(n) = 1, \gcd(n, N) = 1,$$
$$\xi_N(n) = 0, \gcd(n, N) > 1.$$

This is evidently a Dirichlet character mod N, called the **principal character**. It corresponds to the trivial homomorphism $U(N) \to \mathbb{C}^\times$, i.e., the one which maps every element to $1 \in \mathbb{C}$.

Example: $N = 1$: Since $\varphi(1) = 1$, the principal character ξ_1 coincides with the trivial character $\mathbf{1}$: this is the unique Dirichlet character modulo 1.

Example: $N = 2$: Since $\varphi(2) = 1$, the principal character ξ_2, which maps odd numbers to 1 and even integers to 0, is the unique Dirichlet character modulo 2.

Example: $N = 3$: Since $\varphi(3) = 2$, there are two Dirichlet characters mod 3, the principal one ξ_3 and a nonprincipal character, say χ_3. One checks that $\chi_3(n)$ must be 1 if $n = 3k+1$, -1 if $n = 3k+2$, and 0 if n is divisible by 3. Thus $\widehat{U(3)} = \{\xi_3, \chi_3\}$.

Example: $N = 4$: Since $\varphi(4) = 2$, there is exactly one nonprincipal Dirichlet character mod 4, χ_4. We must define $\chi_4(n)$ to be 0 if n is even and $(-1)^{\frac{n-1}{2}}$ if n is odd. Thus $\widehat{U(4)} = \{\xi_4, \chi_4\}$. Note that $\xi_4 = \xi_2$.

[9] In fact, $X(N)$ and $U(N)$ are isomorphic groups: Theorem 15, *ibid.*, but this is actually not needed here.

7.3. Conductors and primitive characters.

7.4. Dirichlet L-series.

By definition, a **Dirichlet L-series** is the Dirchlet series associated to a Dirichlet character:

$$L(\chi, s) = D(\chi, s) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}.$$

In particular, taking $\chi = \chi_1 = \mathbf{1}$, we get $L(\chi_1, s) = \zeta(s)$, which has $\sigma_{ac} = \sigma_c = 1$. But this is the exception:

THEOREM 16.21. *Let χ be a nontrivial Dirichlet character. Then for the Dirichlet L-series $L(\chi, s) = D(\chi, s)$, we have $\sigma_{ac} = 1$, $\sigma_c = 0$.*

PROOF. It follows from the orthogonality relations [Handout A2.5, Theorem 17] that since χ is nonprincipal, the partial sums of $L(\chi, s)$ are bounded. Indeed since $|\chi(n)| \leq 1$ for each n and the sum over any N consecutive values is zero, the partial sums are bounded by N. Also we clearly have $\chi(n) = 1$ for infinitely many n, e.g. for all $n \equiv 1 \pmod{N}$. So the result follows directly from Theorem 16.14. □

We remark that most of the proof of the Dirichlet's theorem – specifically, that every congruence class $n \in (\mathbb{Z}/N\mathbb{Z})^\times$ contains infinitely many primes – involves showing that for every nontrivial character χ, $L(\chi, 1 + it)$ is nonzero for all $t \in \mathbb{R}$. This turns out to be much harder if χ takes on only real values.

8. An Explicit Statement of the Riemann Hypothesis

Let g be the arithmetical function $g(n) = (-1)^{n+1}$. Then:

$$D(g, s) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s} = \sum_{n=1}^{\infty} \frac{1}{n^s} - 2\sum_{n=1}^{\infty} \frac{1}{(2k)^s} = \zeta(s)(1 - 2^{1-s}).$$

This formal manipulation holds analytically on the region on which all series are absolutely convergent, namely on $\Re(s) > 1$. On the other hand, by Example XX above we know that $D(g, s)$ is convergent for $\Re(s) > 0$. So consider the function

$$Z(s) = \frac{D(g, s)}{1 - 2^{1-s}}.$$

By Corollary 16.16 the numerator is complex analytic for $\Re(s) > 0$. The denominator is defined and analytic on the entire complex plane, and is zero when $2^{1-s} = e^{(1-s)\log 2} = 1$, or when $1 - s = \frac{2\pi n i}{\log 2}$ for $n \in \mathbb{Z}$, so when $s = s_n = 1 - \frac{2\pi n}{\log 2} i$. But by construction $Z(s) = \zeta(s)$ for $\Re(s) > 1$, so $Z(s)$ is what is called an **meromorphic continuation** of the zeta function.

Remark: All of the zeroes of 2^{1-s} are simple (i.e., are not also zeroes for the derivative). It follows that for $n \neq 0$, $Z(s)$ is holomorphic at s_n iff $D(g, s_n) = 0$. We will see in the course of the proof of Dirichlet's theorem that this indeed the case, and thus $Z(s) = \zeta(s)$ is analytic in $\Re(s) > 0$ with the single exception of a simple pole at $s = 1$.

However, our above analysis already shows that 2^{1-s} is defined and nonzero in

the *critical strip* $0 < \Re(s) < 1$, so that for such an s, $Z(s) = 0 \iff D(g,s) = 0$. We can therefore give a precise statement of the Riemann hypothesis in the following (misleadingly, of course) innocuous form:

CONJECTURE 16.22. *(Riemann Hypothesis) Suppose s is a zero of the function*

$$D(g,s) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n^s}$$

with $0 < \Re(s) < 1$. Then $\Re(s) = \frac{1}{2}$.

This serves to show once again how the deepest facts (and conjectures!) in analytic number theory turn on cancellation in infinite series.

9. General Dirichlet Series

Let $\lambda = \{\lambda_n\}_{n=1}^{\infty}$ be a sequence of real numbers which is strictly increasing and with $\lim_{n\to\infty} \lambda_n = \infty$. Given a complex sequence (or "arithmetical function") $a = \{a_n\}_{n=1}^{\infty}$, we may consider the series

$$D_\lambda(a,s) = \sum_{n=1}^{\infty} a_n e^{-s\lambda_n},$$

called the **general Dirichlet series** associated to the *sequence of exponents* λ.

The theory we have developed for Dirichlet series can equally well be expressed in this more general context. Why one might want to do this is probably not yet clear, but bear with us for a moment.

In particular, if we define as before σ_{ac} (resp. σ_c) to be the infimum of all real numbers σ such that $\sum_{n=1}^{\infty} |a_n| e^{-\sigma\lambda_n}$ converges (resp. such that $D_\lambda(a,\sigma)$ converges), one can prove that $\Re(s) > \sigma_{ac}$ (resp. $\Re(s) > \sigma$) is the largest open half-plane in which $D_\lambda(a,s)$ is absolutely convergent (resp. convergent). Moreover, there are explicit formulas for these abscissae, at least when $\sigma_c \geq 0$ (which holds in all applications we know of). For instance if $\sum_n a_n$ diverges then $\sigma_c \geq 0$.

THEOREM 16.23. *([**A2**, §8.2]) Let $D_\lambda(a,s)$ be a general Dirichlet series, and assume that $\sigma_c \geq 0$. Then:*

$$\sigma_{ac} = \limsup_n \frac{\log \sum_{k=1}^{n} |a_k|}{\lambda_n}. \tag{53}$$

$$\sigma_c = \limsup_n \frac{\log |\sum_{k=1}^{n} a_k|}{\lambda_n}. \tag{54}$$

Remark: If $\limsup_n \frac{\log |\sum_{k=1}^{n} a_k|}{\lambda_n} = 0$ and $\sum_n a_n$ diverges, then $\sigma_c = 0$;
if $\limsup_n \frac{\log |\sum_{k=1}^{n} a_k|}{\lambda_n} = 0$ and $\sum_n a_n$ converges, then

$$\sigma_c = \limsup_n \frac{1}{\lambda_n} \ln |\sum_{i=1}^{\infty} a_i|.$$

These formulae are highly reminiscent of Hadamard's formula $(\limsup_n |a_n|^{\frac{1}{n}})^{-1}$ for the radius of convergence of a power series $\sum_{n=0}^{\infty} a_n x^n$.

But in fact it is no coincidence: just as general Dirichlet series generalize "ordinary" Dirichlet series – which we recover by taking $\lambda_n = \log n$, they also generalize power series – which we essentially recover by taking $\lambda_n = n$. Indeed,

$$\sum_{n=1}^{\infty} a_n e^{-ns} = \sum_{n=1}^{\infty} a_n x^n,$$

with $x = e^{-s}$. This change of variables takes right half-planes to disks around the origin: indeed the open disk $|x| < R$ corresponds to

$$|x| = |e^{-s}| = |e^{-\sigma - it}| = e^{-\sigma} < R,$$

or $\sigma > -\log R$, a right half-plane. Under the change of variables $x = e^{-s}$ the origin $x = 0$ corresponds to some ideal complex number with infinitely large real part.

At first the fact that we have a theory which simultaneously encompasses Dirichlet series and power series seems hard to believe, since the open disks of convergence and of absolute convergence for a power series are identical. However, the analogue of Proposition 16.12 for general Dirichlet series is

PROPOSITION 16.24. *Let $D_\lambda(a, s)$ be a general Dirichlet series. Then the abscissae of absolute convergence and of convergence are related by:*

$$0 \leq \sigma_{ac} - \sigma_c \leq \limsup_{n \to \infty} \frac{\log n}{\lambda_n}.$$

In the case $\lambda_n = n$ we have $\frac{\log n}{n} \to 0$, and Proposition 16.24 confirms that $\sigma_{ac} = \sigma_c$.

We leave it as an exercise for the interested reader to compare the formulae (53) and (54) with Hadamard's formula $R^{-1} = \limsup_n |a_n|^{\frac{1}{n}}$ for the radius of convergence of power series. (After making the change of variables $x = e^{-s}$ they are not identical formulae, but it is not too hard to show that they are equivalent in the sense that any of them can be derived from the others without too much trouble.)

CHAPTER 17

Dirichlet's Theorem on Primes in Arithmetic Progressions

1. Statement of Dirichlet's Theorem

The aim of this section is to give a complete proof of the following result:

THEOREM 17.1. *(Dirichlet, 1837) Let $a, N \in \mathbb{Z}^+$ be such that $\gcd(a, N) = 1$. Then there are infinitely many prime numbers p such that $p \equiv a \pmod{N}$.*

We remark that the proof gives more, that the set of primes $p \equiv a \pmod{N}$ is **substantial** in the sense of [Handout 12].[1]

One of the amazing things about the proof of Dirichlet's theorem is how modern it feels. It is literally amazing to compare the scope of the proof to the arguments we used to prove some of the other theorems in the course, which historically came much later. Dirichlet's theorem comes 60 years before Minkowski's work on the geometry of numbers and 99 years before the Chevalley-Warning theorem!

Let us be honest that the proof of Dirichlet's theorem is of a difficulty beyond that of anything else we have attempted in this course. On the algebraic side, it requires the theory of characters on the finite abelian groups $U(N) = (\mathbb{Z}/N\mathbb{Z})^\times$. From the perspective of the 21st century mathematics undergraduate with a background in abstract algebra, these are not particularly deep waters. More serious demands come from the analytic side: the main strategy is, as in Euler's proof of the infinitude of primes, to consider the function

$$P_a(s) = \sum_{p \equiv a \pmod{N}} \frac{1}{p^s},$$

which is defined say for real numbers $s > 1$, and to show that $\lim_{s \to 1^+} P_a(s) = +\infty$. Of course this suffices, because a divergent series must have infinitely many terms! The function $P_a(s)$ will in turn be related to a finite linear combination of logarithms of Dirichlet L-series, and the differing behavior of the Dirichlet series for principal and non-principal characters is a key aspect of the proof. Indeed, the fuel for the entire proof is the following surprisingly deep fact:

THEOREM 17.2. *(Dirichlet's Nonvanishing Theorem) For any non-principal Dirichlet character χ of period N, we have $L(\chi, 1) \neq 0$.*

[1] In fact, with relatively little additional work, one can show that the primes are, in a certain precise sense, equidistributed among the $\varphi(N)$ possible congruence classes.

There are many possible routes to Theorem 17.2. We have chosen (following Serre) to present a proof which exploits the theory of Dirichlet series which we have developed in the previous handout in loving detail. As in our treatment of Dirichlet series, we do find it convenient to draw upon a small amount of complex function theory. These result are summarized in Appendix C, which may be most useful for a reader who has not yet been exposed to complex analysis but has a good command of the theory of sequences and series of real functions.

I hope that readers who are unable or unwilling to check carefully through all the analytic details of the proof will still gain an appreciation for the sometimes difficult but also quite beautiful ideas which are on display here. It may be appropriate for me to end this introduction with a personal statement. I believe that I first encountered the proof of Dirichlet's theorem during a reading course in (mostly analytic) number theory that I took as an undergraduate with Professor R. Narasimhan, but in truth I have little memory of it. For my entire graduate career I neglected analysis in general and analytic number theory in particular, to the extent that I came to regard the study of conditionally convergent series as a sort of idle amusement. As a postdoc in Montréal I found myself in an environment where analytic and algebraic number theory were regarded with roughly equal importance (and better yet, often practiced simultaneously). Eventually the limitations of my overly algebraic bias became clear to me, and since my arrival at UGA I have made some progress working my way back towards a more balanced perspective.

Dirichlet's theorem points the way towards modern analytic number theory more than any other single result (even more than the Prime Number Theorem, in my opinion, whose analytic proof is harder but less immediately enlightening). Thus I came to the desire to dicsuss the proof of Dirichlet's theorem in the course (which was not done the first time I taught it).

The proof that I am about to present is not substantively different from what can be found in many other texts (and especially, to the proof given in [**Se73**]). Nevertheless, in order to both follow every detail of the proof and also to get a sense of what was going on in the proof as a whole took me dozens of hours of work, much more so than any other topic in this course. But to finally be able to present the proof feels wonderful, like coming home again. So although I have done what I can to present this material as transparently as possible, not only will I be sympathetic if you find parts of it confusing the first time around, I will even be a little jealous if you don't! But do try to enjoy the ride.

2. The Main Part of the Proof of Dirichlet's Theorem

2.1. Prelude on complex logarithms.

We begin rather inauspiciously by discussing logarithms. By a complex logarithm, we mean a holomorphic function $L(z)$ such that $e^{L(z)} = z$. As compared to the usual real logarithm, there are two subtleties. First, there are multiple such functions: since $e^{z+2\pi in} = e^z$ for all z, if $L(z)$ is any complex logarithm, so is $L(z)+2\pi in$ for any integer n. More seriously, no complex logarithm can be defined on the entire complex plane. Clearly we cannot have a logarithm defined at 0, since 0 is not in the image of the complex exponential function. In complex analysis one learns that if one removes from the complex line any ray passing through the origin –

the real interval $(-\infty, 0]$ being the most standard choice – then one can define a complex logarithm on this restricted domain. In particular, given any open disk in the complex plane which does not contain the origin, there is a complex logarithm defined on that disk.

For the moment though, let us proceed exactly as in calculus: we define a function $\log(1-z)$ for $|z| < 1$ by the following convergent Taylor series expansion:

$$\log(1-z) = -\sum_{n=1}^{\infty} \frac{z^n}{n}. \tag{55}$$

In our analysis, we will come to a point where we have an analytic function, say $f(z)$, and we will want initially want to interpret $\log f(z)$ in a rather formal way, i.e., simply as the series expansion

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

It will be clear for our particular $f(z)$ that the series converges to an analytic function, say g, of z. The subtle point is whether g really is a logarithm of f in the above sense, i.e., whether and for which values of z we have $e^{g(z)} = f(z)$. Our expository choice here is to state carefully the claims we are making about logarithms during the course of the proof and then come back to explain them at the end. Readers with less familiarity with complex analysis may skip these final justifications without fear of losing any essential part of the argument.

2.2. The proof. To begin the proof proper, we let $X(N)$ denote the group of Dirichlet characters modulo N. Fix a with $\gcd(a, N) = 1$ as in the statement of Dirichlet's theorem.

Write $\mathcal{P}_a$ for the set of prime numbers $p \equiv a \pmod{N}$, so our task is of course to show that $\mathcal{P}_a$ is infinite. For this we consider the function

$$P_a(s) := \sum_{p \in \mathcal{P}_a} \frac{1}{p^s},$$

defined for s with $\Re(s) > 1$. Our goal is to show that $P_a(s)$ approaches infinity as s approaches 1. (It would be enough to show this for real σ – i.e., $\lim_{\sigma \to 1^+} P_a(\sigma) = \infty$ – but nevertheless for the proof it is useful to consider complex s.)

Remark: Notice that this gives more than just the infinitude of $\mathcal{P}_a$: it shows that it is "substantial" in the sense of Handout X.X.

The overarching idea of the proof is to express $P_a(s)$ in terms of some Dirichlet L-series for characters $\chi \in X(N)$, and thus to reduce the unboundedness of $P_a(s)$ as $s \to 1$ from some corresponding analytic properties of L-series near $s = 1$.

Why should $P_a(s)$ have anything to do with Dirichlet L-series? First, define $\mathbf{1}_a$ be the characteristic function of the congruence class $a \pmod{N}$: i.e., $\mathbf{1}_a(n)$ is 1 if $n \equiv a \pmod{N}$ and is 0 otherwise. Then $P_a(s)$ is reminiscent of the Dirichlet series for the arithmetical function $\mathbf{1}_a$, except it is a sum only over primes. Note that since $\mathbf{1}_a$ is not a multiplicative function, it would be unfruitful to consider

its Dirichlet series $D(\mathbf{1}_a, s)$ – it does not have an Euler product expansion. Nevertheless $\mathbf{1}_a$ has some character-like properties: it is N-periodic and it is 0 when $\gcd(n, N) > 1$. Therefore $\mathbf{1}_a$ is entirely determined by the corresponding function $U(N) \to C$, $n \pmod N \mapsto \mathbf{1}_a(n)$.

Now recall from [Handout A2.5, §4.3] that any function $f : U(N) \to \mathbb{C}^\times$ can be uniquely expressed as a $\mathbb{C}$-linear combination of characters; [Ibid, Corollary 18] even gives an explicit formula.

With all this in mind, it is easy to discover the following result (which we may as well prove directly):

LEMMA 17.3. *For all $n \in \mathbb{Z}$, we have*

$$\mathbf{1}_a(n) = \sum_{\chi \in X(N)} \frac{\chi(a)^{-1}}{\varphi(N)} \chi(n).$$

Proof: By the complete multiplicativity of the χ's, the right hand side equals

$$\frac{1}{\varphi(N)} \left(\sum_{\chi \in X(N)} \chi(a^{-1}n) \right),$$

and now by orthogonality the parenthesized sum evaluates to $\varphi(N)$ if $a^{-1}n \equiv 1 \pmod N$ – i.e., if $n \equiv a \pmod N$ – and 0 otherwise. The result follows.

The corresponding identity for $P_a(s)$ is:

$$P_a(s) = \sum_{\chi \in X(N)} \frac{\chi(a)^{-1}}{\varphi(N)} \sum_p \frac{\chi(p)}{p^s}. \tag{56}$$

The terms $\sum_p \frac{\chi(p)}{p^s}$ are clearly reminiscent of Dirichlet L-series. Starting with

$$L(\chi, s) = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}.$$

and "taking logarithms" we get

$$\log L(\chi, s) = \sum_p -\log(1 - \frac{\chi(p)}{p^s}).$$

Expanding out this logarithm using the series (55) as advertised above, we get

$$\log(L(\chi, s)) = \sum_p \sum_n (\frac{\chi(p)}{p^s})^n / n. \tag{57}$$

But we regard the above as just being "motivational"; let us now be a little more precise. The right hand side of (57) is absolutely convergent for $\Re(s) > 1$ and uniformly convergent on closed half-planes $\Re(s) \geq 1 + \delta$. So if we simply *define*

$$\ell(\chi, s) := \sum_p \sum_n \left(\frac{\chi(p)}{p^s}\right)^n / n,$$

then, whatever else it may be, $\ell(\chi, s)$ is an analytic function on the half-plane $\Re(s) > 1$. Of course we know what the "whatever else" should be:

First Claim on Logarithms: In the halfplane $\Re(s) > 1$, we have $e^{\ell(\chi,s)} = L(\chi, s)$.

As stated above, we postpone justification of this claim until the next section.

Notice that the $n = 1$ contribution to $\ell(\chi, s)$ alone gives precisely the sums appearing in (56); there are also all the $n \geq 2$ terms, which we don't want. So let's separate out these two parts of the series: definining

$$\ell_1(\chi, s) = \sum_p \frac{\chi(p)}{p^s}$$

and

$$R(\chi, s) = \sum_{n \geq 2} \sum_p \frac{\chi(p)^n}{np^{ns}},$$

we have

$$\ell(\chi, s) = \ell_1(\chi, s) + R(\chi, s)$$

and also

$$P_a(s) = \sum_{\chi \in X(N)} \frac{\chi(a^{-1})}{\varphi(N)} \ell_1(\chi, s).$$

But recall what we're trying to show: that $P_a(s)$ is unbounded as $s \to 1$. If we're trying to show that something is bounded, any terms which *do* remain bounded as $s \to 1$ can be ignored. But

$$|R(\chi, 1)| \leq \sum_{n \geq 2} \sum_p \frac{1}{np^n} \leq \sum_p \sum_{n \geq 2} \left(\frac{1}{p}\right)^n$$

$$= \sum_p \frac{1}{p^2} \frac{p}{p-1} \leq \sum_p \frac{1}{p^2} \cdot 2 \leq 2 \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

So $R(\chi, s)$ is absolutely convergent at $s = 1$ hence remains bounded as $s \to 1$, and thus we can safely ignore the terms $R(\chi, s)$. The following notation expresses this:

$$P_a(s) = \sum_{\chi \in X(N)} \frac{\chi(a)^{-1}}{\varphi(N)} \ell(\chi, s) + O(1);$$

here the "$O(1)$" denotes anything which is uniformly bounded as $s \to 1$. Separating the term corresponding to the principal character ξ_N from the other terms, we get

$$P_a(s) = \frac{1}{\varphi(N)} \sum_{p \nmid N} p^{-s} + \sum_{\chi \neq \xi_N} \ell(\chi, s) + O(1).$$

Now $\sum_{p \nmid N} p^{-s}$, is up to a finite number of terms, just the sum $\sum_p p^{-s}$. We know well that $\sum_p p^{-1} = \infty$, and by the Positivity Lemma this implies $\lim_{s \to 1^+} \sum_p p^{-s} = +\infty$. So the first term is unbounded near infinity. Therefore it would suffice to show that for each *nontrivial* character χ, $\ell(\chi, s)$ is bounded as $s \to 1$.

Recall that for every nonprincipal χ we know that the Dirichlet series for $L(\chi, s)$ is convergent on all of $\Re(s) > 0$; in particular, $L(\chi, s)$ is a well-defined analytic

function at $s = 1$. Finally, we see the relevance of Theorem 17.2: if we know that for each nonprincipal $\chi \in X(N)$, $L(\chi, 1) \neq 0$, then

$$L(\chi, 1) = \lim_{s \to 1} L(\chi, s) = \lim_{s \to 1} e^{\ell(\chi, s)}.$$

Second claim on logarithms: Therefore $\ell(\chi, s)$ is bounded as $s \to 1$.

Now, modulo these two claims and the proof of Theorem 17.2 we're done: since the contribution to $P_a(s)$ from the nonprincipal Dirichlet L-series remains bounded as $s \to 1$ whereas the contribution from the principal Dirichlet L-series does not, it follows that $P_a(s)$ itself is unbounded as s approaches 1: more precisely, as s approaches 1 through real values of $s > 1$, we get

$$\lim_{s \to 1^+} P_a(s) = \sum_{p \equiv a \pmod N} \frac{1}{p^{-s}} = +\infty,$$

hence there must be infinitely many primes $p \equiv a \pmod N$.

2.3. Tidying up the logarithms.

Let us now deal with our two claims on logarithms. For the first one, we know from calculus that for a real number s with $|s| < 1$, the Taylor series expansion

$$-\log(1 - s) = \sum_{n=1}^{\infty} \frac{s^n}{n}$$

is valid: in other words, we do have the identity

$$e^{-\sum_{n=1}^{\infty} \frac{s^n}{n}} = 1 - s$$

for all such s. By the principle of analytic continuation, the corresponding complex power series gives a well-defined logarithm whenever it is defined, which is at least for complex s with $|s| < 1$. We have

$$\lim_{\Re(s) \to +\infty} L(\chi, s) = 1,$$

so that there exists a σ_0 such that $\Re(s) > \sigma_0$ implies $|1 - L(\chi, s)| < 1$. Thus in this halfplane we do have $e^{\ell(\chi, s)} = L(\chi, s)$. By the principle of analytic continuation, this identity will continue to hold so long as both sides are well-defined analytic functions, which is the case for all $\Re(s) > 1$, justifying the first claim on logarithms.

Similar reasoning establishes the second claim: since $L(\chi, s)$ is analytic and nonzero at $s = 1$, there exists some small open disk about $L(\chi, 1)$ which does not contain the origin, and therefore we can choose a branch of the logarithm such that $\log L(\chi, s)$ is well-defined on the preimage of that disk, so in particular on some small open disk D about $s = 1$. Then $\log L(\chi, 1)$ is a well-defined complex number. It may not be equal to our $\ell(\chi, 1)$, but since any two logarithms of the same analytic function differ by a constant integer multiple of $2\pi i$, by the principle of analytic continuation there exists some $n \in \mathbb{Z}$ such that $\ell(\chi, s) - 2\pi n = \log(\chi, s)$ on the disk D, and no matter what n is, this means that $\ell(\chi, s)$ remains bounded as $s \to 1$.

3. Nonvanishing of $L(\chi, 1)$

We claim that $L(\chi, 1) \neq 0$ for all nonprincipal characters $\chi \in X(N)$. Our argument is as follows: consider the behavior of the Dedekind zeta function

$$\zeta_N(s) = \prod_{\chi \in X(N)} L(\chi, s).$$

near $s = 1$. We know that for each nonprincipal χ, $L(\chi, s)$ is holomorphic at $s = 1$, whereas for principal χ we get essentially the Riemann zeta function, which we have seen has a simple pole at $s = 1$: we have seen that

$$(s-1)\zeta(s) \to 1$$

as $s \to 1$. It follows from basic function theory that $\zeta_N(s)$ has at most a simple pole at $s = 1$, and indeed has a pole iff $L(\chi, 1) \neq 0$ for all nontrivial χ. Thus our goal is to show that the Dedekind zeta function $\zeta_N(s)$ has a singularity at $s = 1$.

The key is that the Dirichlet series $\zeta_N(s)$ has a very particular form. To see this, we need just a little notation: for a prime p not dividing N, let $f(p)$ denote the order of p in the unit group $U(N)$, and put $g(p) = \frac{\varphi(N)}{f(p)}$, which is by Lagrange's theorem a positive integer. Now:

PROPOSITION 17.4. *a) We have*

$$\zeta_N(s) = \prod_{p \nmid N} \frac{1}{\left(1 - \frac{1}{p^{f(p)s}}\right)^{g(p)}}.$$

b) Therefore $\zeta_N(s)$ is a Dirichlet series with non-negative integral coefficients, converging absolutely in the half-plane $\Re(s) > 1$.

PROOF. Let $\mu_{f(p)}$ be the group of $f(p)$th roots of unity. Then for all $p \nmid N$ we have the polynomial identity

$$\prod_{w \in \mu_{f(p)}} (1 - wT) = 1 - T^{f(p)}.$$

Indeed, both sides have the $f(p)$th roots of unity as roots (with multiplicity one), so they differ at most by a multiplicative constant; but both sides evaluate to 1 at $T = 0$. Now by the Character Extension Lemma [Lemma 13, Handout A2.5], for all $w \in \mu_{f(p)}$ there are precisely $g(p)$ elements $\chi \in X(N)$ such that $\chi(p) = w$. This establishes part a), and part b) follows from the explicit formula of part a). □

Now for a *deus ex machina*. We are given that $\zeta_N(s)$ is a Dirichlet series with non-negative real coefficients. Therefore we can apply Landau's Theorem: if σ is the abscissca of convergence of the Dirichlet series, then the function $\zeta_N(s)$ has a singulariety at σ. Clearly $\sigma \geq 1$, so, contrapositively, if $\zeta_N(s)$ does not have a singularity at $s = 1$, then not only does $\zeta_N(s)$ extend analytically to some larger halfplane $\Re(s) > 1 - \epsilon$, but it extends until it meets a singularity on the real line. But we have already seen that each Dirichlet L-series is holomorphic for $0 < \Re(s) < 1$, so Landau's theorem tells us that $\sigma \leq 0$.

If you think about it for a minute, it is exceedingly unlikely that a Dirichlet series with non-negative integral coefficients has absissca of convergence $\sigma \leq 0$, and in

our case it is quite straightforward to see that this is not the case: take s to be in the real interval $(0,1)$. Expanding out the pth Euler factor we get

$$\frac{1}{\left(1-\frac{1}{p^{f(p)s}}\right)^{g(p)}} = \left(1+\frac{1}{p^{f(p)s}}+\frac{1}{p^{2f(p)s}}+\dots\right).$$

Ignoring all the crossterms gives a crude upper bound: this quantity is at least

$$1+\frac{1}{p^{\varphi(N)s}}+\frac{1}{p^{2\varphi(N)s}}+\dots.$$

Multiplying this over all p, it follows that

$$\zeta_N(s) \geq \sum_{n \mid (n,N)=1} \frac{1}{n^{\varphi(N)s}}.$$

When we evaluate at $s=\frac{1}{\varphi(N)}$ we get

$$\sum_{(n,N)=1} \frac{1}{n}.$$

Since the set of integers prime to N has positive density, it is substantial. More concretely, since every n of the form $Nk+1$ is coprime to N, this last sum is at least as large as

$$\sum_{k=1}^{\infty} \frac{1}{Nk+1} = \infty.$$

QED!

CHAPTER 18

Rational Quadratic Forms and the Local-Global Principle

A **form** of degree k is a polynomial $P(x_1, \ldots, x_n)$ which is homogeneous of degree k: in each monomial term $cx_1^{i_1} \cdots x_n^{i_n}$, the *total degree* $i_1 + \ldots + i_n$ is k. E.g.

$$F_n(x, y, z) = x^n + y^n - z^n$$

is a form of degree n, such that the study of solutions to $F_n(x, y, z) = 0$ is equivalent to Fermat's Last Theorem.

For the most part we will concentrate here on **quadratic forms** ($k = 2$):

$$\sum_{1 \leq i \leq j \leq n} a_{ij} x^i x^j,$$

where the coefficients a_{ij} are usually either integers or rational numbers (although we shall also be interested in quadratic forms with coefficients in $\mathbb{Z}/n\mathbb{Z}$ and $\mathbb{R}$). For instance, a **binary** quadratic form is any expression of the form

$$q(x, y) = ax^2 + bxy + cy^2.$$

As for most Diophantine equations, quadratic forms were first studied over the integers, meaning that the coefficients a_{ij} are integers and only integer values of $x_1, \ldots, x_n$ are allowed to be plugged in. At the end of the 19th century it was realized that by allowing the variables $x_1, \ldots, x_n$ to take *rational* values, one gets a much more satisfactory theory. (In fact one can study quadratic forms with coefficients and values in any field F. This point of view was developed by Witt in the 1930's, expanded in the middle years of this century by, among others, Pfister and Milnor, and has in the last decade become especially closely linked to one of the deepest and most abstract branches of contemporary mathematics: "homotopy K-theory.") However, a wide array of firepower has been constructed over the years to deal with the complications presented by the integral case, culminating recently in some spectacular results. Here we will concentrate on what can be done over the rational numbers, and also on what statements about integral quadratic forms can be directly deduced from the theory of rational quadratic forms.

Let us distinguish two types of problems concerning a quadratic form $q(x_1, \ldots, x_n)$, which we will allow to have *either* integral or rational coefficients a_{ij}.

Homogeneous problem (or **isotropy problem**): Determine whether there exist integers, $x_1, \ldots, x_n$, not all zero, such that $q(x_1, \ldots, x_n) = 0$. A quadratic form such that $q(x) = 0$ has a nontrivial integral solution is said to be **isotropic**; if there is no nontrivial solution it is said to be **anisotropic**.

Example 0: The sum of squares forms $x_1^2 + \ldots + x_n^2$ are all anisotropic. Indeed, for any real numbers $x_1, \ldots, x_n$, not all zero, $x_1^2 + \ldots + x_n^2 > 0$: a form with this property is said to be **positive definite**.

Example 1: The $\mathbb{Z}$-quadratic form $x^2 - ny^2$ is isotropic iff n is a perfect square.

Inhomogeneous problem: For a given integer n, determine whether the equation $q(x_1, \ldots, x_n) = n$ has an integer solution (if so, we say "q **represents** n"). More generally, for fixed q, determine all integers n represented by q.

Example 2: We determined all integers n represented by a $x_1^2 + x_2^2$, and stated without proof the results for the quadratic forms $x_1^2+x_2^2+x_3^2$ and $x_1^2+x_2^2+x_3^2+x_4^2$; in the latter case, all positive integers are represented.

In general the inhomogeneous problem is substantially more difficult than the homogeneous problem. One reason why the homogeneous problem is easier is that, even if we originally state it in terms of the integers, it can be solved using rational numbers instead:

PROPOSITION 18.1. *(Principle of homogeneous equivalence) Let $P(x_1, \ldots, x_n)$ be a homogeneous polynomial with integral coefficients. Then $P(x_1, \ldots, x_n)$ has a nontrivial solution with $x_1, \ldots, x_n \in \mathbb{Z}$ iff it has a nontrivial solution with $x_1, \ldots, x_N \in \mathbb{Q}$.*

PROOF. Of course a nontrivial integral solution is in particular a nontrivial rational solution. For the converse, assume there exist $\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}$, not all 0, such that $P(\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}) = 0$. Suppose P is homogeneous of degree k. Then for any $\alpha \in \mathbb{R}^\times$, we have

$$P(\alpha x_1, \ldots, \alpha x_n) = \alpha^k P(x_1, \ldots, x_n),$$

since we can factor out k α's from every term. So let $N = \mathrm{lcm}(q_1, \ldots, q_n)$. Then

$$P(N\frac{p_1}{q_1}, \ldots, N\frac{p_n}{q_n}) = N^k P(\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}) = N^k \cdot 0 = 0,$$

so that $(N\frac{p_1}{q_1}, \ldots, N\frac{p_n}{q_n})$ is a nontrivial integral solution. □

Thus the homogeneous problem for integral forms (of any degree) is really a problem about *rational* forms.

Remark: The inhomogeneous problem still makes sense for forms of higher degree, but to solve it – even for rational forms – is generally extremely difficult. For instance, Selmer conjectured in 1951 that a prime $p \equiv 4,\ 7,\ 8 \pmod 9$ is of the form $x^3 + y^3$ for two *rational* numbers x and y. A proof of this in the first two cases was announced (but not published) by Noam Elkies in 1994; more recently, Dasgupta and Voight have carefully written down a proof of a slightly weaker result [**DV09**]. The case of $p \equiv 8 \pmod 9$ remains open. In this case (i.e., that of binary cubic forms) the rich theory of rational points on elliptic curves can be fruitfully applied. Even less is known about (say) binary forms of higher degree.

1. Rational Quadratic Forms

In this section, we work with quadratic forms q with coefficients a_{ij} lying in $\mathbb{Q}$. (In fact, everything we say works over an arbitrary field F whose characteristic is different from 2.) This gives many advantages, which we state mostly without proof:

Fact 1: Every rational quadratic form can be diagonalized.

In general, two quadratic forms q and q' should be regarded as **equivalent** if there is an invertible linear change of variables $(x_1', \ldots, x_n') = A(x_1, \ldots, x_n)$ carrying one to the other. In particular, equivalent quadratic forms represent the same values, and equivalence preserve an/isotropy.

Any quadratic form $q(x_1, \ldots, x_n)$ can be represented by a symmetric matrix Q, such that

$$q(x_1, \ldots, x_n) = xQx^T,$$

where $x = (x_1, \ldots, x_n)$. However, there is a slight annoyance here which is seen by calculating the quadratic form associated to the symmetric matrix

$$\begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

; it is

$$q(x_1, x_2) = ax_1^2 + 2bx_1x_2 + bx_2^2$$

So to get the "general" binary quadratic form of XX, we need to use the matrix

$$\begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & d, \end{bmatrix}$$

and in general, the symmetric matrix M corresponding to the quadratic form $\sum_{i \leq j} a_{ij} X^i X^j$ is

$m_{ij} = a_{ij}$, $i = j$,
$m_{ij} = \frac{a_{ij}}{2}$, $i \neq j$,

so the representing matrix M of an integral quadratic form q will in general have only half-integral entries.

Now the matrix interpretation of equivalence is as follows: the form with representing matrix M is equivalent to the quadratic form with representing matrix AMA^T for any invertible matrix A. If we are working with rational quadratic forms, then M and A can have rational entries and the condition for invertibility is that $\det(A) \neq 0$. However, if we are working with integral quadratic forms, then A must have integral entries and its inverse must have integral entries, which means that $\det(A) = \pm 1$.

Recall from linear algebra that every real symmetric matrix M is similar to a diagonal matrix via a matrix A which is orthogonal: $A^{-1} = A^T$. In fact, for every symmetric matrix M with entries in a subfield F of $\mathbb{C}$, there exists an invertible matrix A such that AMA^T is diagonal: this amounts to saying that we can "rationally diagonalize" a symmetric matrix by performing simultaneous row and column operations. We omit the proof.

In particular, every rational quadratic form is equivalent to a quadratic form of the shape

$$\langle a_1, \ldots, a_n \rangle = a_1 x_1^2 + \ldots + a_n x_n^2.$$

Example: Consider the integral quadratic form $q(x, y) = xy$, with associated matrix $M = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$. Note that we have $\det(M) = -\frac{1}{4}$. If there exists an integrally invertible matrix A with $AMA^T = D$ diagonal, then

$$\det(D) = \det(A)\det(M)\det(A^T) = \det(M)\det(A)^2 = \det(M) = -\frac{1}{4}.$$

But the diagonal entries of the matrix defining an integral quadratic form must be integers, so the determinant of any integrally diagonalizable quadratic form must be an integer. So $q(x, y) = xy$ is not integrally diagonalizable.

Fact 2: Every isotropic rational quadratic form is universal.

There is a special quadratic form

$$H = \langle 1, -1 \rangle = x_1^2 - x_2^2,$$

called the **hyperbolic plane**. By diagonalizing the form $q(x, y) = xy$, one sees that it is equivalent, over $\mathbb{Q}$, to H. In particular the hyperbolic plane H is isotropic – indeed take $x_1 = x_2$ – and moreover it represents every nonzero scalar $x \in \mathbb{Q}^\times$: take $y = 1$. One can show that if q is any isotropic rational quadratic form, then

$$q \cong x_1^2 - x_2^2 + q'(x_3, \ldots, x_n),$$

so that every isotropic form "contains" the hyperbolic plane. In particular, every quadratic form which is isotropic *rationally* represents every rational number.

This is not true over $\mathbb{Z}$: the isotropic quadratic form $x^2 - y^2$ does not represent every integer. Indeed, $x^2 - y^2 \equiv 2 \pmod 4$ has no solution, so $x^2 - y^2$ does not represent any integer which is $2 \pmod 4$.

Fact 3: Over $\mathbb{Q}$, the representation problem can be reduced to the isotropy problem.

More precisely, one has the following result:

THEOREM 18.2. *Let $q(x_1, \ldots, x_n)$ be a quadratic form over $\mathbb{Q}$ (or over any field F of characteristic different from 2), and let $a \in \mathbb{Q}^\times$ (or $a \in F^\times$). The following are equivalent:*
a) The quadratic form $q(x_1, \ldots, x_n) + (-a)x_{n+1}^2$ is isotropic.
b) The quadratic form q rationally represents a.

PROOF. If q represents a, then there exist $x_1, \ldots, x_n \in \mathbb{Q}$, not all 0, such that $q(x_1, \ldots, x_n) = a$, but then rewriting gives

$$q(x_1, \ldots, x_n) + (-a)(1)^2 = 0.$$

Conversely, suppose there are $x_1, \ldots, x_n, x_{n+1}$ in $\mathbb{Q}$, not all 0, such that $q(x_1, \ldots, x_n) + (-a)x_{n+1}^2 = 0$. If $x_{n+1} \neq 0$, then we can move it to the other side and divide by it

(thank goodness we are over $\mathbb{Q}$!) to get

$$q(\frac{x_1}{x_{n+1}}, \ldots, \frac{x_n}{x_{n+1}}) = a.$$

Otherwise, we have $q(x_1, \ldots, x_n) = 0$, for $x_1, \ldots, x_n$ not all zero, which means that q is isotropic, and we averred above that this implies that q "contains" the hyperbolic plane H and therefore represents *every* element of $\mathbb{Q}^\times$, in particular a. □

Thus, if we had an algorithm for deciding whether a given rational quadratic form is isotropic, then applying it to the form $q + (-a)x_{n+1}^2$, we can equally well decide whether it rationally represents any given number a.

Remark: There is, to the best of my knowledge, absolutely nothing like "Fact 3" for forms of higher degree.

2. Legendre's Theorem

We can now give a complete solution to a problem we first considered early on in the course: given $a, b, c \in \mathbb{Z}$, how do we know whether the (quadratic) form

$$ax^2 + by^2 + cz^2 = 0$$

has a nontrivial solution?

Note that, by the discussion of the last section, if we can solve this problem we can completely solve the homogeneous problem for ternary *integral* quadratic forms $q(x, y, z) = 0$. Indeed, by Proposition 18.1 it is enough to decide whether or not $q(x, y, z) = 0$ has a nontrivial rational solution, and working rationally we can diagonalize q to get an equation of the above form.

The answer is given by the following beautiful theorem of Legendre. To state it, we will employ some *ad hoc* notation: for nonzero integers a and b, we will write $a\square\ b$ to mean that a is a square (possibly zero) modulo $|b|$. Note that, if b is odd, $a\square\ b$ implies that the Jacobi symbol $\left(\frac{a}{b}\right) = 1$, but not conversely. A small lemma:

LEMMA 18.3. *Let $b, c \in \mathbb{Z} \setminus \{0\}$, with $\gcd(b, c) = 1$. Then $a\square\ bc \iff a\square\ b$ and $a\square\ c$.*

If there exists an integer x such that $a \equiv x^2 \pmod{bc}$, then certainly $a \equiv x^2 \pmod{b}$ and $a \equiv x^2 \pmod{c}$, giving the forward implication. Conversely, if $a \equiv x^2 \pmod{b}$ and $a \equiv y^2 \pmod{c}$, then since b and c are relatively prime, by CRT there exists a $z \pmod{bc}$ such that $z \equiv x \pmod{b}$ and $z \equiv y \pmod{c}$, hence $a \equiv z^2 \pmod{b}$ and $a \equiv z^2 \pmod{c}$, so $a \equiv z^2 \pmod{bc}$.

THEOREM 18.4. *(Legendre) Let a, b, c be nonzero integers, squarefree, relatively prime in pairs, and neither all positive nor all negative. Then*

$$ax^2 + by^2 + cz^2 = 0$$

has a nontrivial integral solution iff all of the following hold:
(i) $-ab\square\ c$.
(ii) $-bc\square\ a$.
(iii) $-ca\square\ b$.

Some remarks on the conditions: if a, b and c are all positive or all negative, the quadratic form is definite over $\mathbb{R}$ and has no nontrivial real solutions. Because integral isotropy is equivalent to rational isotropy, we may adjust a, b and c by any rational square, and therefore we may assume that they are squarefree integers. Moreover, if two of them are divisible by a prime p, then they are both exactly divisible by p, and by a simple ord_p argument the equation certainly has no solutions unless p divides c. But then we may divide through a, b and c by p.

Let us prove the easy half of this theorem now, namely showing that these conditions are necessary. In fact, let us show that they are precisely the conditions obtained by postulating a primitive integral solution (x, y, z) and going modulo a, b and c. Indeed, go modulo c: we get

$$ax^2 \equiv -by^2 \pmod{c};$$

multiplying by $-b$, which is coprime to c, we get the equivalent condition

$$-abx^2 \equiv (by)^2 \pmod{c}.$$

Suppose first that there exists some prime $p \mid c$ such that $p \mid x$. Then since $\gcd(b, c) = 1$, we get $p \mid y$, and that implies $p^2 \mid -ax^2 - by^2 = cz^2$. Since c is squarefree, this implies $p \mid z$, contradicting primitivity. Therefore x is nonzero modulo every prime p dividing c, so x is a unit modulo c, and we can divide, getting

$$-ab \equiv (byx^{-1})^2 \pmod{c},$$

which is condition (i). By symmetry, reducing modulo a we get (ii) and reducing modulo b we get (iii).

Following Ireland and Rosen, to prove the sufficiency we will state the theorem in an equivalent form, as follows:

THEOREM 18.5. *(Legendre's theorem restated) For a and b positive squarefree integers, the equation*

$$ax^2 + by^2 = z^2$$

has a nontrivial integral solution iff all of the following hold:
(i) $a\square\, b$.
(ii) $b\square\, a$.
(iii) $-\frac{ab}{d^2}\square\, d$, where $d = \gcd(a, b)$.

We leave it as a (not difficult, but somewhat tedious) exercise to the reader to check that Theorem 18.5 is equivalent to Theorem 18.4.

Now we prove the sufficiency of the conditions of Theorem 18.5.

The result is obvious if $a = 1$.

Case 1: $a = b$. The theorem asserts that $ax^2 + ay^2 = z^2$ has a solution iff -1 is a square modulo a. By the first supplement to QR, this is last condition is equivalent to: no prime $p \equiv 3 \pmod 4$ divides a. If this condition holds then by the two squares theorem we have $a = r^2+s^2$, and then we can take $x = r,\ y = s,\ z = r^2+s^2$. On the other hand, if there exists $p \mid a$, $p \equiv 3 \pmod 4$, then taking ord_p of both sides of the equation $z^2 = a(x^2 + y^2)$ gives a contradiction, since $\mathrm{ord}_p(z^2) = 2\,\mathrm{ord}_p(z)$ is

even, and $\mathrm{ord}_p(a(x^2+y^2)) = \mathrm{ord}_p(a) + \mathrm{ord}_p(x^2+y^2) = 1 + \mathrm{ord}_p(x^2+y^2)$ implies $\mathrm{ord}_p(x^2+y^2)$ is odd, contradicting the Two Squares Theeorem.

If $b > a$, we can interchange a and b, so we may now assume that $a > b$.

We will now prove the theorem by a descent-type argument, as follows: assuming the hypotheses of Theorem 18.5 we will construct a new form $Ax^2+by^2 = z^2$ satisfying the same hypotheses, with $0 < A < a$, and such that if this latter equation has a nontrivial solution then so does $ax^2+by^2 = z^2$. We perform this reduction process repeatedly, interchanging A and b if $A < b$. Since each step reduces $\max(A,b)$, eventually we will be in the case $A = 1$ or $A = b$, in which we have just shown the equation has a solution. Reversing our sequence of reductions shows that the original equation has a solution.

Now, since $b\square\ a$, there exist T and c such that

$$c^2 - b = aT, \tag{58}$$

for $T \in \mathbb{Z}$. Applying the square/squarefree decomposition, we may write $T = Am^2$ with A squarefree. Choosing c minimally, we may assume that $|c| \leq \frac{a}{2}$.

Claim: $0 < A < a$.

Proof: Since $0 \leq c^2 = aAm^2 + b < a(Am^2+1)$ and $a > 0$, $Am^2 > -1$; since b is squarefree, $T = am^2 \neq 0$, hence $Am^2 \geq 1$ and thus $A > 0$. Also

$$aAm^2 < c^2 \leq \frac{a^2}{4},$$

so

$$A \leq Am^2 < \frac{a}{4} < a.$$

Claim: $A\square\ b$.

Recalling $d = \gcd(a,b)$, write $a = a_1 d$, $b = b_1 d$, so that $\gcd(a_1, b_1) = 1$; since a and b are squarefree, this implies $\gcd(a_1, d) = \gcd(b_1, d) = 1$. Then (58) reads

$$c^2 - b_1 d = a_1 d A m^2 = aAm^2.$$

So $d \mid c^2$, and since d is squarefree, $d \mid c$. Put $c = c_1 d$ and cancel:

$$dc_1^2 - b_1 = Aa_1 m^2. \tag{59}$$

So $Aa_1m^2 \equiv -b_1 \pmod{d}$; multiplying through by a_1, we get

$$Aa_1^2 m^2 \equiv -a_1 b_1 \pmod{d}. \tag{60}$$

Now, any common prime factor p of m and d would divide both b_1 and d, a contradiction; so $\gcd(m,d) = 1$. Since $\frac{-ab}{d^2} = -a_1 b_1$ is a square modulo d by (iii) and a_1 and m are units modulo d, (60) implies that $A\square\ d$. Moreover, $c^2 \equiv aAm^2 \pmod{b_1}$. Since $a\square\ b$, $a\square\ b_1$. Also $\gcd(a, b_1) = 1$ – a common divisor would divide d, but $\gcd(b_1, d) = 1$ – and similarly $\gcd(m, b_1) = 1$. So

$$A \equiv c^2 (am^2)^{-1} \pmod{b_1},$$

and hence $A\square\ b_1$. Since $A\square\ b_1$ and $A\square\ d$, by Lemma 18.3 $A\square\ b$.

Next, put $r = \gcd(A, b)$ and $A = rA_1$, $b = rb_2$, so that $\gcd(A_1, b_2) = \gcd(r, b_2) = 1$. We claim that $-A_1 b_2 \square\, r$. Using (58) we have

$$c^2 - rb_2 = c^2 - rb = aAm^2 = arA_1 m^2. \tag{61}$$

Since b is squarefree, so is r, hence $r \mid c$. So if a prime p divides both am and r, then $p^2 \mid c^2 - aAm^2 = rb_2 \implies p^2 \mid r$, a contradiction. So $\gcd(am, r) = 1$. Putting $c = rc_1$,

$$arA_1 m^2 \equiv -rb_2 \pmod{r^2},$$

so

$$aA_1 m^2 \equiv -b_2 \pmod{r}.$$

Since $a \square\, b$ and $r \mid b$, $a \square\, r$. Multiplying through by b_2, we get

$$-aA_1 b_2 m^2 \equiv b_2^2 \pmod{r},$$

and since $\gcd(am, r) = 1$, we conclude $-A_1 b_2 \square\, r$.

Now assume that $AX^2 + bY^2 = Z^2$ has a nontrivial solution. Then

$$AX^2 = Z^2 - bY^2. \tag{62}$$

Multiplying (62) by (58), we have

$$a(AXm)^2 = (Z^2 - bY^2)(c^2 - b) = (Zc + bY)^2 - b(cY + Z)^2.$$

Note that this unlikely-looking identity can be interpreted as

$$N(Z + Y\sqrt{b})N(c + \sqrt{b}) = N(Zc + bY + (cY + Z)\sqrt{b}).$$

Putting $x = AXm$, $y = cY + Z$, $z = Zc + bY$, this gives a solution to $ax^2 + by^2 = z^2$, which is nontrivial since $x \neq 0$. Thus we have completed our "descent" argument, which proves that the equation has a solution.

3. Hilbert's Reciprocity Law

As we mentioned, Legendre's theorem has the following consequence: a ternary quadratic form

$$q_{a,b} : aX^2 + bY^2 - Z^2$$

has a nontrivial integral solution iff there is a real solution and for every prime p and every positive integer a the congruence

$$aX^2 + bY^2 \equiv Z^2 \pmod{p^a} \tag{63}$$

has a nontrivial solution. As a increases, each of these congruences is stronger than the last, so it makes some sense to bundle up the infinitely many questions of whether any of these p-power congruences has a solution into a single question. Let us introduce the following terminology:

An integral quadratic form $q(x_1, \ldots, x_n)$ is **p-isotropic** if for all $a \in \mathbb{Z}^+$, the congruence $q(x_1, \ldots, x_n) \equiv 0 \pmod{p^a}$ has a nontrivial solution. Otherwise we will say that it is **p-anisotropic**. We will say that q is **∞-isotropic** if it has a real solution.[1]

Considering the case of $q_{a,b}$, for each prime p and for ∞ we are asking a yes/no

[1]Don't ask why we have introduced the symbol ∞ to describe the real solutions. It is just traditional to do so. Moreover, we will eventually get tired of saying "(and ∞)" and start writing $p \leq \infty$. There is no need to read anything deep into this, at least not today.

question – "Is $q_{a,b}$ p-isotropic?" so it makes some vague sense to denote "yes" by $+1$ and "no" by -1, so we define **symbols** $\langle a,b\rangle_p$ for all primes p and $\langle a,b\rangle_\infty$ in this way: i.e., $+1$ if $q_{a,b}$ is p-isotropic and -1 if it is p-anisotropic (and the same for ∞). So Legendre's theorem can be rephrased by saying that $q_{a,b}$ is istropic iff $\langle a,b\rangle_p = 1$ for all $p \leq \infty$.

But now that we've defined the notation, a further question occurs to us: if $q_{a,b}$ is isotropic, the answers to our questions are always yes; but if it isn't, at least some of the answers are no. So which combinations of yes and no are possible?

THEOREM 18.6. *(Hilbert) a) For every pair of nonzero integers a, b, the symbol $\langle a,b\rangle_p$ is equal to $+1$ except possibly for finitely many values of $p \leq \infty$.*
b) Two integral ternary quadratic forms $q_{a,b}$ and $q_{c,d}$ are rationally equivalent *– i.e., one can be obtained from the other by a 3×3 invertible matrix A with rational entries – iff $\langle a,b\rangle_p = \langle c,d\rangle_p$ for all $p \leq \infty$.*
c) The finite set of $p \leq \infty$ for which $\langle a,b\rangle_p = -1$ has an even *number of elements.*
d) For every subset S of the primes union ∞ which is finite and of even order, there exist a and b such that $\langle a,b\rangle_p = -1$ iff $p \in S$.

We admit that this is a mouthful. In particular parts b) and d) solve yet a third problem on rational quadratic forms: their *classification* up to equivalence. We advise the reader to concentrate on the following consequence: for any $q_{a,b}$, by part a) we can consider the infinite product $\prod_{p\leq\infty}\langle a,b\rangle_p$ (since it equals 1 except possible finitely many times), and by part c) we get the following relation, the **Hilbert reciprocity law**:

$$\prod_{p\leq\infty}\langle a,b\rangle_p = 1 \tag{64}$$

This has the extremely useful upshot that instead of having to check congruences modulo all powers of all primes and a sign condition, it suffices to omit any one $p \leq \infty$ from these checks. In particular, we could omit "$p = \infty$" from the checking and get the following result which looks hard to believe based upon the proof we gave: if $ax^2 + by^2 = z^2$ has a solution modulo p^a for all p and a, then it necessarily has an integral solution: in particular the condition that a and b are not both positive *follows automatically* from all the congruence conditions, although it is certainly independent of any finite number of them!

In fact, with a bit of hindsight one can see that the condition of whether or not there is going to be a solution modulo all powers of 2 is the most complicated one. This is taken into account in the statement of Legendre's theorem: the congruence conditions on their own would not imply that $\langle a,b\rangle_2 = +1$ without the sign conditions ("conditions at ∞"), so somehow Legendre's clean conditions exploit this slight redundancy. To see this, consider the case of $a = b = -1$, which has solutions modulo every power of an odd prime, but no nontrivial solutions modulo 4 (and also no real solutions).

Hilbert also found explicit formulae for $\langle a,b\rangle_p$ in terms of Legendre symbols. For the sake of concision we do not state it here. However, we cannot help but mentioning that if one knows these formulae (which are not so hard to prove), then the relation (64) is equivalent to knowing quadratic reciprocity together with its first and second supplements! It turns out that all aspects of the theory rational

quadratic forms can be generalized to the case where the coefficients lie not in $\mathbb{Q}$ but in an arbitrary algebraic number field K. In particular, a suitable version of Hilbert's reciprocity law holds over K, and this is a very clean way to phrase quadratic reciprocity over number fields.

4. The Local-Global Principle

We are now in a position to state what is surely one of the most important and influential results in all of number theory.

THEOREM 18.7. *(Hasse-Minkowski) Let $q(x_1, \ldots, x_n)$ be an integral quadratic form. The following are equivalent:*
a) q is isotropic (over $\mathbb{Z}$ $\iff$ over $\mathbb{Q}$).
b) q is isotropic over $\mathbb{R}$, and for all $n \in \mathbb{Z}^+$, there are nontrivial solutions to the congruence $q(x_1, \ldots, x_n) \equiv 0 \pmod{n}$.

It is clear that a) $\implies$ b). Indeed, in contrapositive form, this has been our favorite "easy" method for showing that an equation *does not* have a solution: any integral solution also gives a real solution and a solution to every possible congruence. The matter of it is in the converse, which asserts that if a quadratic form $q(x_1, \ldots, x_n) = 0$ does not have an integral solution, we can always detect it via congruences and/or over the real numbers.

This turns out to be the master theorem in the area of rational quadratic forms. It is not (yet) stated in a form as explicit as Legendre's theorem for ternary quadratic forms – which, recall, did not just assert that isotropy modulo n for all n implied isotropy over $\mathbb{Z}$ (or equivalently, over $\mathbb{Q}$) but actually said explicitly, in terms of the coefficients, a finite list of congruence conditions to check. Indeed one knows such explicit conditions in all cases, and we will return to mention them in the next section, but for now let us take a broader approach.

First, even in its "qualitative form" the theorem gives an algorithm for determining whether any quadratic form is isotropic. Namely, we just have to search in parallel for one of the two things:

(i) Integers $x_1, \ldots, x_n$, not all 0, such that $q(x_1, \ldots, x_n) = 0$.
(ii) An integer N such that the congruence $q(x_1, \ldots, x_n) \equiv 0 \pmod{N}$ has only the all-zero solution.

For any given N, (ii) is a finite problem: we have exactly $N^n - 1$ values to plug in and see whether we get 0. Similarly, if we wanted to check all tuples of integers $(x_1, \ldots, x_n)$ with $\max_i |x_i| \leq M$, then that too is obviously a finite problem. Conceivably we could search forever and never find either a value of M as in (i) or a value of N as in (ii) – for sure we will never find both! – but the Hasse-Minkowski Theorem asserts that if we search long enough we will find either one or the other. This then is our algorithm!

In point of fact the situation is better for part (ii): it can be shown that for any degree k form $P(x_1, \ldots, x_n)$ with integer coefficients, there is a recipe (algorithm!) for computing a *single* value of N such that if $P(x_1, \ldots, x_n) \equiv 0 \pmod{N}$ has a

nontrivial solution, then *for all* N the congruence has a solution. Moreover, one can determine whether or not there are any real solutions (using methods from calculus). For this the two essential tools are:

(i) The Weil bounds for points on curves over $\mathbb{Z}/p\mathbb{Z}$, which allows one to compute a finite set of primes S such that for all $p > S$ the congruence $P \equiv 0 \pmod{p}$ automatically has nontrivial solutions (in fact, a number of solutions which tends to ∞ with p).

This is a serious piece of mathematics dating from around the 1940's.

(ii) Hensel's Lemma, which gives sufficient conditions for lifting a solution $(x_1, \ldots, x_n)$ to $P \equiv 0 \pmod{p}$ to solutions modulo all higher powers p^a of p.

This turns out to be surprisingly similar to Newton's method for finding roots of equations, and the proof is relatively elementary.

Alas, we do not have time to say more about either one.

So in finite time we can determine whether or not there is any value of N for which $P(x_1, \ldots, x_n) \equiv 0$ has only the trivial solution, and we can also tell whether there are real solutions. Of course, if $P = 0$ fails to have congruential solutions and/or real solutions, then we know it cannot have nontrivial integral (equivalently, rational) solutions. But suppose we find that our form P passes all these tests? Can we then assert that it has a nontrivial integral solution?

As we have just seen (or heard), the answer is a resounding "yes" when P is a **quadratic** form. In general, whenever the answer to this question is "yes", one says that the **local-global principle**, or **Hasse principle**, holds for P. Of course the big question is: does the Hasse principle hold for all forms of higher degree?

One can also ask whether the Hasse principle holds for not-necessarily homogeneous polynomials, like $x^2 + y^3 + z^7 = 13$. The following remarkable result shows that it could not possibly hold for all polynomials in several variables over the integers.

THEOREM 18.8. *(Davis-Matijasevic-Putnam-Robinson) There is no algorithm that will accept as input a polynomial $P(x_1, \ldots, x_n)$ with integral coefficients and output 1 if $P(x_1, \ldots, x_n) = 0$ has an integral solution, and 0 otherwise.*

Since we just said that there is an algorithm which determines if a polynomial (not necessarily homogeneous, in fact) has congruential solutions and real solutions, there must therefore be some polynomials which pass these tests and yet still have no solutions.

Remark: It is unknown whether there exists an algorithm to decide if a polynomial with rational coefficients has a rational solution.

One might think that such counterexamples to the Hasse principle might be in some sense nonconstructive, but this is not at all the case:

THEOREM 18.9. *The following equations have congruential solutions and real solutions, but no nontrivial integral solutions:*
a) (Selmer) $3X^3 + 4Y^3 + 5Z^3 = 0$*;*
b) (Bremner) $5w^3 + 9x^3 + 10y^3 + 12z^3 = 0$.

These are just especially nice examples. It is known (if not "well-known") that for every $k > 2$ there is a form $P(x, y, z) = 0$ of degree k which violates the local-global principle. In fact some of my own work has been devoted to constructing large (in particular, infinite) sets of counterexamples to the local-global principle.

There are however some further positive results, the most famous and important being the following:

THEOREM 18.10. *(Birch) Let k be a positive integer. Then there exists an $n_0(k)$ with the following property:*
a) If k is odd, then every degree k form $P(x_1, \ldots, x_n) = 0$ in $n \geq n_0$ variables has a nontrivial integral solution.
b) If k is odd and $P(x_1, \ldots, x_n)$ is a degree k form in $n \geq n_0$ variables with "low-dimensional singularities", then P has a nontrivial integral solution iff it has a nontrivial real solution.

Remark: The condition of low-dimensional singularities is a bit technical. Let us rather define what it means for an equation to have no singularities at all, which is a special case. A nontrivial **complex** solution $(x_1, \ldots, x_n)$ to $P(x_1, \ldots, x_n)$ at which all the partial derivatives $\frac{\partial P}{\partial x_i}$ vanish is called a **singular point**. (Perhaps you remember from multivariable calculus these are the points at which a curve or surface can be "not so nice": i.e., have self-intersections, cusps, or other pathologies.) P is said to be **nonsingular** if there are no singular points. In particular, one immediately checks that a diagonal form $P(x_1, \ldots, x_n) = a_1 x_1^k + \ldots + a_k x_n^k$ is nonsingular, so Birch's theorem applies to diagonal forms, and in particular to quadratic forms. (As far as I know it is an open problem whether the theorem holds for forms of even degree without any additional hypotheses.)

Thus morally, if only there are enough variables compared to the degree, then all congruence conditions are automatically satisfied and moreover th. However, in the proof n_0 does indeed have to be very large compared to k, and it is quite an active branch of analytic number theory to improve upon these bounds.

Another idea, which we shall be able to express only vaguely and see an example of in the case of the inhomogeneous problem for integral quadratic forms, is that if one asks as a yes/no question whether or not the existence of congruential solutions and real solutions is enough to ensure the existence of integral solutions, then one has to take rather drastic measures – e.g., enormously many variables compared to the degree, as above – to ensure that the answer is "yes" rather than "no" most of the time. However, if one can somehow **quantify** the failure of a local-global phenomenon, then one can hope that in any given situation it fails only to a *finite* extent.

5. Local Conditions for Isotropy of Quadratic Forms

(ii) Although the result is not phrased in explicit form, part of the point is that one can easily determine whether the condition of part b) holds. For instance, there will be real solutions unless, when the quadratic form is diagonalized (over $\mathbb{Q}$), all of the diagonal entries have the same sign. It is less obvious but still true that given *any* equation $P(x_1, \ldots, x_n)$, there is an algorithm to check in a finite amount of time whether for *all* N, $P(x_1, \ldots, x_n) \equiv 0 \pmod{N}$ has nontrivial solutions. Explicit conditions will be given in the case of ternary quadratic forms ($n = 3$), coming up soon. Such conditions are known for all n (for $n = 2$, they are the restrictions coming from quadratic reciprocity that we have already seen).

(iii) In fact as the number of variables increases it becomes much easier to satisfy the congruence conditions, until we get to $n = 5$: every quadratic form $q(x_1, \ldots, x_n)$ in 5 or more variables has nontrivial solutions modulo every integer N! This has a remarkable corollary:

THEOREM 18.11.
a) Let $q(x_1, \ldots, x_n)$ be an integral quadratic form in at least 5 variables. Then $q(x) = 0$ has a nontrivial integral solution iff it has a nontrivial real solution*, i.e., unless q is positive or negative definite.*
b) Let q be a quadratic form in at least 4 variables which is not negative (resp. positive) definite – i.e., over $\mathbb{R}$ it takes on some positive (resp. negative) values. Then q rationally represents all positive (resp. negative) rational numbers.

PROOF. Part a) follows immediately from the Hasse-Minkowski theorem and the assertion that there are no "congruential" obstructions to a quadratic form in at least 5 variables being isotropic. Part b) follows morally by applying Theorem 18.2, although to see it one needs to know that there is a field $\mathbb{Q}_p$ of characteristic 0 with the property that q is isotropic over $\mathbb{Q}_p$ iff q is isotropic modulo p^a for all a. □

We deduce in particular that every positive rational number is a sum of four rational squares. This is of course weaker than Lagrange's Theorem, and it must be, because the theorem also applies e.g. to $2x_1^2 + 3x_2^2 + 4x_3^2 + 5x_4^2$, which visibly does not represent 1 over $\mathbb{Z}$.

CHAPTER 19

Representations of Integers by Quadratic Forms

As we have seen, if

$$P(x_1, \ldots, x_n) = d$$

is an inhomogeneous polynomial equation (i.e., $d \neq 0$), then the determination of whether it has an integer solution is considerably more subtle than whether it has a rational solution. Perhaps the best single example of this is the proven nonexistence of an algorithm to determine whether a polynomial equation has an integral solution. In contrast, the question of whether a homogeneous polynomial equation must have a nontrivial solution is equivalent to the issue of whether polynomial equations must have rational solutions, and this is a wide open problem (although some experts think that it too will turn out to be algorithmically undecidable).

We have just surveyed the complete theory of homogeneous quadratic equations in any number of variables. One of the great miracles of the quadratic case is that, over $\mathbb{Q}$, the inhomogeneous problem reduces to the homogeneous problem, so that given a quadratic form $q(x_1, \ldots, x_n)$, we now know how to determine the set of all integers (or even rational numbers) d such that

$$q(x_1, \ldots, x_n) = d$$

has a *rational* solution. Two of the more striking consequences we derived from this Hasse-Minkowski theory were the following:

Fact 1: A quaternary quadratic form $q = ax_1^2 + bx_2^2 + cx_3^2 + dx_4^2$ rationally represents all integers allowed by sign considerations:
(i) if a, b, c, d are all positive, q represents all $d \in \mathbb{Q}^{>0}$;
(ii) if a, b, c, d are all negative, q represents all $d \in \mathbb{Q}^{<0}$;
(iii) otherwise q represents all $d \in \mathbb{Q}^{\times}$.

Fact 2: The three squares form $x^2 + y^2 + z^2$ rationally represents an integer d iff $d > 0$ and $d \neq 4^a(8k + 7)$.

These are strongly reminiscent of two results we stated but not did prove for integral quadratic forms, namely that $x_1^2 + x_2^2 + x_3^2 + x_4^2$ *integrally* represents all positive integers and $x_1^2 + x_2^2 + x_3^2 + x_4^2$ *integrally* represents all positive integers except precisely those of the form $4^a(8k + 7)$.

It seems clear that we cannot hope to recover general integral representability results from the Hasse-Minkowski theory. For instance, Fact 1 does not distinguish between the Four Squares form and a form in which a, b, c, d are all at least 2: such a form clearly cannot represent 1 integrally! Morally speaking, "local conditions"

of congruence and sign do not take into account the *size* of the coefficients of the quadratic form, whereas one clearly wants some or all of the coefficients to be small in order for a positive definite quadratic form to have a fighting chance at representing small positive integers.

So what to do?

Let us describe some of the ways that various mathematicians have reacted to this question over the years.

1. The Davenport-Cassels Lemma

Here is a beautiful observation which allows us to solve the representation problem for $x^2 + y^2 + z^2$:

LEMMA 19.1. *(Davenport-Cassels) Let $q(x) = f(x_1, \ldots, x_n) = \sum_{i,j=1}^n a_{ij} x_i x_j$ be a quadratic form with $a_{ij} = a_{ji} \in \mathbb{Z}$. We suppose **condition (DC)**: that for any $y = (y_1, \ldots, y_n) \in \mathbb{Q}^n \setminus \mathbb{Z}^n$, there exists $x = (x_1, \ldots, x_n) \in \mathbb{Z}^n$ such that*

$$0 < |q(x - y)| < 1.$$

Then, for any integer d, q represents d rationally iff q represents d integrally.

PROOF. For $x, y \in \mathbb{Q}^n$, put $x \cdot y := \frac{1}{2}(q(x+y) - q(x) - q(x))$. Then $(x, y) \mapsto x \cdot y$ is bilinear and $x \cdot x = q(x)$. Note that for $x, y \in \mathbb{Z}^n$, we need not have $x \cdot y \in \mathbb{Z}$, but certainly we have $2(x \cdot y) \in \mathbb{Z}$. Our computations below are parenthesized so as to emphasize this integrality property.
Let $d \in \mathbb{Z}$, and suppose that there exists $x \in \mathbb{Q}^n$ such that $q(x) = d$. Equivalently, there exists $t \in \mathbb{Z}$ and $x' \in \mathbb{Z}^n$ such that $t^2 d = x' \cdot x'$. We choose x' and t such that $|t|$ is minimal, and it is enough to show that $|t| = 1$.

Applying the hypothesis (DC) $x = \frac{x'}{d}$, there exists a $y \in \mathbb{Z}^n$ such that if $z = x - y$ we have

$$0 < |q(z)| < 1.$$

Now put

$$a = y \cdot y - d,$$
$$b = 2(dt - x' \cdot y),$$
$$T = at + b,$$
$$X = ax' + by.$$

Then $a, b, T \in \mathbb{Z}$, and $X \in \mathbb{Z}^n$.
CLAIM: $X \cdot X = T^2 d$.
Indeed,

$$X \cdot X = a^2(x' \cdot x') + ab(2x' \cdot y) = b^2(y \cdot y) = a^2 t^2 d + ab(2dt - b) + b^2(d + a)$$
$$= d(a^2 t^2 + 2abt + b^2) = T^2 d.$$

CLAIM: $T = t(z \cdot z)$.
Indeed,

$$tT = at^2 + bt = t^2(y \cdot y) - dt^2 + 2dt^2 - t(2x' \cdot y)$$
$$= t^2(y \cdot y) - t(2x' \cdot y) + x' \cdot x' = (ty - x') \cdot (ty - x') = (-tz) \cdot (-tz) = t^2(z \cdot z).$$

Since $0 < |z \cdot z| < 1$, we have $0 < |T| < |t|$, contradicting the minimality of $|t|$. $\square$

Remark 1: Suppose that the quadratic form q is anisotropic. Then condition (DC) is equivalent to the following more easily verified one: for all $x \in \mathbb{Q}^n$, there exists $y \in \mathbb{Z}^n$ such that $|q(x-y)| < 1$. Indeed, since $x \notin \mathbb{Z}^n$ and $y \in \mathbb{Z}^n$, $x - y \notin \mathbb{Z}^n$. In particular $x - y \neq (0, \ldots, 0)$, so since q is anistropic, necessarily $|q(x-y)| > 0$.

Remark 2: Lemma 19.1 has a curious history. So far as I know there is no paper of Davenport and Cassels (two eminent 20th century number theorists) which contains it: it is more folkloric. The attribution of this result seems to be due to J.-P. Serre in his influential text [**Se73**]. Later, André Weil pointed out [**W**] that in the special case of $f(x) = x_1^2 + x_2^2 + x_3^2$, the result goes back to a 1912 paper of the amateur mathematician L. Aubry [**Au12**].

There is also more than the usual amount of variation in the hypotheses of this result. Serre's text makes the additional hypothesis that f is positive definite – i.e., $x \neq 0 \implies f(x) > 0$. Many of the authors of more recent number theory texts that include this result follow Serre and include the hypothesis of positive definiteness. Indeed, when I first wrote these notes in 2006, I did so myself (and included a place-holder remark that I belived that this hypothesis was superfluous).[1] To get from Serre's proof to ours requires only (i) inserting absolute values where appropriate, and (ii) noting that whenever we need $x \cdot y$ to be integral, we have an extra factor of 2 in the expression to make it so. The result is also stated and proved (in a mildly different way) in Weil's text.

Remark 3: In the isotropic case, the stronger hypothesis $0 < |q(x-y)| < 1$ is truly necessary. Consider for instance $q(x,y) = x^2 - y^2$: we ask the reader to show that 2 is represented rationally but not integrally.

One might call a quadratic form **Euclidean** if it satisfies (DC). For example, the quadratic form $q(x,y) = x^2 - dy^2$ is Euclidean iff given rational numbers r_x, r_y, we can find integers n_x, n_y such that

$$|(r_x - n_x)^2 - d(r_y - n_y)^2| < 1 \tag{65}$$

Since we know that we can find an integer within $\frac{1}{2}$ of any rational number (and that this estimate is best possible!), the quantity in question is at most $(\frac{1}{2})^2 + |d|(\frac{1}{2})$ if $d < 0$ and at most $\frac{d}{4}$ when $d > 0$. So the values of d for which (65) holds are precisely $d = -1, -2, 2, 3$. This should be a familiar list: these are precisely the values of d for which you proved that $\mathbb{Z}[\sqrt{d}]$ is a PID. Whenever $\mathbb{Z}[\sqrt{d}]$ is a PID, one can use Euclid's Lemma to solve the problem of which primes (and in fact which integers, with more care) are integrally represented by $x^2 - dy^2$. The Davenport-Cassels Lemma allows for a slightly different approach: for these values of d, $x^2 - dy^2 = N$ has an integral solution iff it has a rational solution iff $x^2 - dy^2 - Nz^2 = 0$ is isotropic, which we can answer using Legendre's Theorem.

Also $x^2 + y^2 + z^2$ satisfies the hypotheses of the Davenport-Cassels lemma: given rational numbers x, y, z, find integers n_1, n_2, n_3 at most $\frac{1}{2}$ a unit away, and then

$$(x - n_1)^2 + (x - n_2)^2 + (x - n_3)^2 \leq \frac{1}{4} + \frac{1}{4} + \frac{1}{4} < 1.$$

[1] A notable exception is Lam's 2005 text on quadratic forms, which states the result for anisotropic forms, simplified as in Remark 1.

2. The Three Squares Theorem

Our goal in this section is to prove the following celebrated result.

THEOREM 19.2. *(Legendre-Gauss) For $n \in \mathbb{Z}^+$, the following are equivalent:*
(i) n is not *of the form $4^a(8k+7)$ for any $a \in \mathbb{N}$ and $k \in \mathbb{Z}$.*
(ii) n is a sum of three integer squares: there are $x, y, z \in \mathbb{Z}$ with $x^2+y^2+z^2 = n$.

2.1. Proof of the Three Squares Theorem.

The strategy of proof is as follows: the quadratic form $q(x,y,z) = x^2+y^2+z^2$ satisfies the hypotheses of the Davenport-Cassels Lemma (Lemma 19.1) of the previous section. Therefore, to show that an integer n is a sum of three integer squares it suffices to show the *a priori* much weaker assertion that it is a sum of three rational squares. It is traditional to establish the latter assertion using the Hasse-Minkowski theory of quadratic forms over $\mathbb{Q}$ in terms of quadratic forms over the p-adic numbers. But since in these notes we have not even officially introduced the p-adic numbers, we need to do something more elementary. Instead we follow the second half of a short and clever argument of J. Wójcik [**Wó72**], which succeeds in replacing the Hasse-Minkowski Theory with an appeal to (i) Fermat's Two Squares Theorem, (ii) Legendre's Theorem on homogeneous ternary quadratic equations and (iii) Dirichlet's Theorem on Primes in Arithmetic Progressions.[2]

Let us first dispose of the (easy!) direction (i) $\implies$ (ii) of Theorem 19.2.

LEMMA 19.3. *Let n be an integer of the form $4^a(8k+7)$ for some $a \in \mathbb{N}$, $k \in \mathbb{Z}$. Then n is* not *the sum of three rational squares.*

PROOF. Step 0: Suppose on the contrary that $4^a(8k+7)$ is a sum of three rational squares. We may take our rational numbers to have a common deminator $d > 0$ and thus

$$\left(\frac{x}{d}\right)^2 + \left(\frac{y}{d}\right)^2 + \left(\frac{z}{d}\right)^2 = 4^a(8k+7).$$

Clearing denominators, we get

$$x^2+y^2+z^2 = d^2 4^a(8k+7).$$

Write $d = 2^b d'$ with d' odd. Since $1^2, 3^2, 5^2, 7^2 \equiv 1 \pmod 8$, we find that $d'^2 \equiv 1 \pmod 8$ and thus

$$d^2 4^a(8k+7) = (2^b)^2(d'^2)4^a(8k+7) = 4^{a+b}(8k'+7).$$

In other words, to show that no integer of the form $4^a(8k+7)$ is a sum of 3 rational squares, it suffices to show that no integer of the form $4^a(8k+7)$ is a sum of three integral squares. So let us now show this.
Step 1: We observe that $x^2+y^2+z^2 \equiv 7 \pmod 8$ has no solutions. Indeed, since the squares mod 8 are $0, 1, 4$, this is a quick mental calculation. (In particular this disposes of the $a = 0$ case.)
Step 2: we observe that if $n \equiv 0, 4 \pmod 8$ then the congruence

$$x^2+y^2+z^2 \equiv n \pmod 8$$

[2]That we have given complete proofs of all of these theorems previously is a happy coincidence: I did not learn about Wójcik's argument until 2011, more than four years after these notes were first written.

has no *primitive solutions*, i.e., no solutions in which at least one of x, y, z is odd. Indeed, since the squares mod 8 are $0, 1, 4$, so in particular the only odd square is 1. Since 4 and 0 are both even, if x, y, z are not all even, then exactly one two of them must be odd, say x and y, so $x^2 \equiv y^2 \equiv 1 \pmod 8$ and thus $z^2 \equiv 4 - 2 \pmod 8$ or $z^2 \equiv 8 - 2 \pmod 8$, and neither 2 nor 6 is a square modulo 8.
Step 3: Now suppose that there are integers x, y, z such that $x^2+y^2+z^2 = 4^a(8k+7)$. If $a = 0$ then by Step 1 reducing modulo 8 gives a contradiction. If $a = 1$, then $4^a(8k + 7) \equiv 4 \pmod 8$, so by Step 2 any representation $x^2 + y^2 + z^2 = 4(8k + 7)$ must have x, y, z all even, and then dividing by 4 gives $(\frac{x}{2})^2+(\frac{y}{2})^2+(\frac{z}{2})^2 = (8k+7)$, a contradiction. If $a \geq 2$, then $4^a(8k + 7) \equiv 0 \pmod 8$, and again by Step 2 in any representation $x^2 + y^2 + z^2 = 4^a(8k + 7)$ we must have x, y, z all even. Thus writing $x = 2X$, $y = 2Y$, $z = 2Z$ we get an integer representation $X^2 + Y^2 + Z^2 = 4^{a-1}(8k+7)$. We may continue in this way until we get a representation of $4(8k+7)$ as a sum of three integral squares, which we have just seen is impossible. □

LEMMA 19.4. *Suppose that every squarefree positive integer $n \not\equiv 7 \pmod 8$ is a sum of three integral squares. Then every positive integer $n \neq 4^a(8k + 7)$ is a sum of three integral squares.*

PROOF. Let n be a positive integer which is *not* of the form $4^a(8k+7)$. As for any positive integer, we may write n as $n = 2^a n_1^2 n_2$, where $a \geq 0$, n_1 is odd and n_2 is odd and squarefree.
Case 1: $0 \leq a \leq 1$, $n_2 \not\equiv 7 \pmod 8$. Then $2^a n_2$ is squarefree and not 7 (mod 8), so by assumption there exist $x, y, z \in \mathbb{Z}$ such that $x^2 + y^2 + z^2 = 2^a n_2$, and thus $(n_1x)^2 + (n_1y)^2 + (n_1z)^2 = 2^a n_1^2 n_2 = n$.
Case 2: $n_2 \not\equiv 7 \pmod 8$. In such a case n is of the form $(2^b)^2$ times an integer n of the type considered in Case 1. Since such an integer n is a sum of three integreal squares, so is any square times n.
Case 3: $n_2 \equiv 7 \pmod 8$. For n not to be of the form $4^a(8k + 7)$, the power of a must be odd; in other words, we may write n as a square times $2n_2$ where n_2 is squarefree and of the form $8k_+7$. Thus $2n_2$ is squarefree and not of the form $8k+7$, so by assumption $2n_2$ is a sum of three squares, hence so is n. □

LEMMA 19.5. *Let $m \in \mathbb{Z}^+$, $n \equiv 3 \pmod 8$, and write $m = p_1 \cdots p_r$. Then the number of i such that $p_i \equiv 3, 5 \pmod 8$ is even.*

Exercise: Prove Lemma 19.5. (Suggestion: use the Jacobi symbol $\left(\frac{-2}{m}\right)$.)

Since $x^2 + y^2 + z^2$ rationally represents an integer n iff it integrally represents an integer n, the following result completes the proof of Theorem 19.2.

PROPOSITION 19.6. *Let n be a squarefree integer, $n \not\equiv 7 \pmod 8$. Then n is a sum of three rational squares.*

PROOF. To fix ideas we will first give the argument under certain additional congruence conditions and then explain how to modify it to deal with the other cases. Filling in the details for these latters cases would be a good exercise for the interested reader.
Case 1: Let us suppose that $m = p_1 \cdots p_r$ is squarefree and $m \equiv 1 \pmod 4$. Thus each p_i is odd and the number of $p_i \equiv 3 \pmod 4$ is even. By Dirichlet's Theorem on Primes in Arithmetic Progressions, there is a prime number q such that

• $\left(\frac{q}{p_i}\right) = \left(\frac{-1}{p_i}\right)$ for all $1 \leq i \leq p_i$ and
• $q \equiv 1 \pmod 4$.
(Indeed, each of the first conditions restricts q to a nonempty set of congruence classes modulo the distinct odd primes p_i, whereas the last condition is a condition modulo a power of 2. By the Chinese Remainder Theorem this amounts to a set of congruence conditions modulo $4p_1 \cdots p_r$ and all of the resulting congruence classes are relatively prime to $4p_1 \cdots p_r$, so Dirichlet's Theorem applies.)
It follows that for all $1 \leq i \leq r$,

$$\left(\frac{-q}{p_i}\right) = \left(\frac{-1}{p_i}\right)\left(\frac{q}{p_i}\right) = 1,$$

and

$$\left(\frac{m}{q}\right) = \left(\frac{p_1}{q}\right) \cdots \left(\frac{p_r}{q}\right) = \left(\frac{q}{p_1}\right) \cdots \left(\frac{q}{p_r}\right) = \left(\frac{-1}{p_1}\right) \cdots \left(\frac{-1}{p_r}\right) = 1.$$

The last equality holds because the number of factors of -1 is the number of primes $p_i \equiv 3 \pmod 4$, which as observed above is an even number.
since $-q$ is a square modulo each of the distinct primes p_i, by the Chinese Remainder Theorem it is also a square modulo $m = p_1 \cdots p_r$. Therefore by the Chinese Remainder Theorem there is an integer x such that

$$x^2 \equiv -q \pmod m$$

$$x^2 \equiv m \pmod q.$$

But according to Legendre's Theorem, these are precisely the congruence conditions necessary and sufficient for the homogeneous equation

$$qu^2 + z^2 - mt^2 = 0$$

to have a solution in integers (u, z, t), not all zero. Indeed, we must have $t \neq 0$, for otherwise $qu^2 + z^2 = 0 \implies u = z = 0$. Moreover, since $q \equiv 1 \pmod 4$, by Fermat's Two Squares Theorem there are $x, y \in \mathbb{Z}$ such that $qu^2 = x^2 + y^2$. Therefore

$$mt^2 - z^2 = qu^2 = x^2 + y^2,$$

so

$$m = \left(\frac{x}{t}\right)^2 + \left(\frac{y}{t}\right)^2 + \left(\frac{z}{t}\right)^2$$

and m is a sum of three rational squares, completing the proof in this case.
Case 2: Suppose $m = 2m_1 = 2p_1 \cdots p_r$ with $m_1 = p_1 \cdots p_r$ squarefree and odd. In this case we may proceed exactly as above, except that we require $q \equiv 1 \pmod 8$.
Case 3: Suppose $m = p_1 \cdots p_r$ is squarefree and $m \equiv 3 \pmod 8$. By Lemma 19.5, the number of prime divisors p_i of m which are either 5 or 7 modulo 8 is even. By Dirichlet's Theorem there exists a prime q such that
• $\left(\frac{q}{p_i}\right) = \left(\frac{-2}{p_i}\right)$ for all $1 \leq i \leq p_i$ and
• $q \equiv 5 \pmod 8$.
It follows that for all $1 \leq i \leq r$,

$$\left(\frac{-2q}{p_i}\right) = \left(\frac{-2}{p_i}\right)\left(\frac{q}{p_i}\right) = 1,$$

and

$$\left(\frac{m}{q}\right) = \left(\frac{p_1}{q}\right)\cdots\left(\frac{p_r}{q}\right) = \left(\frac{q}{p_1}\right)\cdots\left(\frac{q}{p_r}\right) = \left(\frac{-2}{p_1}\right)\cdots\left(\frac{-2}{p_r}\right) = 1.$$

The last equality holds because the number of factors of -1 is the number of primes $p_i \equiv 5, 7 \pmod 8$, which as observed above is an even number.
Therefore there is an integer x such that

$$x^2 \equiv -2q \pmod m$$
$$x^2 \equiv m \pmod q,$$

so by Legendre's Theorem the equation

$$2qu^2 + z^2 - mt^2 = 0$$

has a solution in integers (u, z, t) with $t \neq 0$. Since $q \equiv 1 \pmod 4$, there are $x, y \in \mathbb{Z}$ such that $2qu^2 = x^2 + y^2$, so

$$mt^2 - z^2 = 2qu^2 = x^2 + y^2,$$

and thus once again

$$m = \left(\frac{x}{t}\right)^2 + \left(\frac{y}{t}\right)^2 + \left(\frac{z}{t}\right)^2.$$

□

2.2. Some applications of the Three Squares Theorem.

Knowing exactly which integers are represented by $x^2 + y^2 + z^2$ turns out to be a powerful weapon for analyzing representation of integers by certain quaternary quadratic forms.

PROPOSITION 19.7. *The three squares theorem implies the four squares theorem.*

PROOF. In order to show the Four Squares Theorem it suffices to show that every squarefree positive integer m is a sum of four integer squares. By the Three Squares Theorem, m is even a sum of three integer squares unless $m = 8k+7$. But if $m = 8k + 7$, then $m - 1 = 8k + 6$. Now $\mathrm{ord}_2(8k + 6) = 1$, so $8k + 6$ is not of the form $4^a(8k+7)$, hence $8k+6 = x^2+y^2+z^2$ and $m = 8k+7 = x^2+y^2+z^2+1^2$. □

More generally:

THEOREM 19.8. *For any $1 \leq d \leq 7$, the quadratic form $q = x^2 + y^2 + z^2 + dw^2$ integrally represents all positive integers.*

PROOF. As above it is enough to show that q represents all squarefree positive integers. Moreover, if $m \neq 8k + 7$ is a squarefree positive integer then m is represented already by $x^2+y^2+z^2$ so certainly by q. It remains to dispose of $m = 8k+7$.
Case 1: Suppose $d = 1, 2, 4, 6$. Then $m - d \cdot 1^2 = m - d$ is:
• $m - 1 = 8k + 6$, if $d = 1$. This is a sum of 3 squares.
• $m - 2 = 8k + 5$, if $d = 2$. This is a sum of 3 squares.
• $m - 4 = 8k + 3$, if $d = 3$. This is a sum of 3 squares.
• $m - 5 = 8k + 2$, if $d = 5$. This is a sum of 3 squares.
• $m - 6 = 8k + 1$, if $d = 6$. This is a sum of 3 squares.
Case 2: If $d = 3$, then

$$m - d \cdot 2^2 = m - 12 = 8k - 5 = 8(k - 1) + 3.$$

Thus, so long as $m-12$ is positive, it is a sum of three squares. We need to check separately that positive integers less than 12 are still represented by q, but this is easy: the only one which is not already a sum of 3 squares is $7 = 2^2+0^2+0^2+3\cdot 1^2$.
Case 3: If $d = 7$, then

$$m - d \cdot 2^2 = m - 28 = 8(k-3) + 5.$$

Thus, so long as $m-28$ is positive, it is a sum of three squares. Again we must separately check that positive integers less than 28 are represented by q, and again this comes down to checking 7: $7 = 0^2 + 0^2 + 0^2 + 7 \cdot 1^2$. □

If we are looking for quaternary quadratic forms $q = x^2 + y^2 + z^2 + dw^2$ which represent *all* positive integers, then we have just found all of them: if $d > 7$, then such a q cannot integrally represent 7. Nevertheless we can still use the Gauss-Legendre Theorem to analyze these forms. For instance.

PROPOSITION 19.9. *For a positive integer n, TFAE:*
(i) There are integers x, y, z, w such that $n = x^2 + y^2 + z^2 + 8w^2$.
(ii) $n \not\equiv 7 \pmod 8$.

PROOF. (i) $\implies$ (ii): For any integers x, y, z, w, reducing $n = x^2+y^2+z^2+8w^2$ modulo 8 gives $n \equiv x^2 + y^2 + z^2 \pmod 8$, and we already know that this has no solutions when $n \equiv 7 \pmod 8$.
(ii) $\implies$ (i): Write $n = 2^a m$ with m odd. If m is not of the form $8k+7$ then both m and $2m$ are sums of three integer squares, and since n is an even power of 2 times either m or $2m$, n must be a sum of three intege squares. So we are reduced to the case $n = 2^a(8k+7)$ with $a \geq 1$. If $a = 1$ then $\mathrm{ord}_2(n) = 1$ and again n is a sum of three integer squares. Suppose $a = 2$, so $n = 32k + 28$ and thus $n - 8 \cdot 1^2 = 32k + 20 = 4(8k+5)$ is of the form $x^2 + y^2 + z^2$ and thus $n = x^2 + y^2 + z^2 + 8w^2$. If $a \geq 3$ is odd, then n is a sum of three squares. If $a \geq 4$ is even, then $n = (2^{\frac{a-2}{2}})^2(4 \cdot (8k+7))$ is a square times an integer represented by q, so n is also represented by q. □

Exercise: Prove or disprove the following claims:
a) If d is a positive integer which is not divisible by 8, then the quadratic form $x^2 + y^2 + z^2 + dw^2$ integrally represents all sufficiently large positive integers.
v) If $d = 8d'$ is a positive integer, then the quadratic form $x^2 + y^2 + z^2 + dw^2$ integrally represents all sufficiently large positive integers which are *not* $7 \pmod 8$.

3. Approximate Local-Global Principle

From now on we restrict to the case of positive-definite integral quadratic forms $q(x_1, \ldots, x_n)$. For such a form, the equation

$$q(x_1, \ldots, x_n) = N$$

can have at most finitely many integral solutions. Indeed, if we define $r_q(N)$ to be the number of solutions, then the summatory function

$$R_q(N) = \sum_{i=1}^{N} r_q(i)$$

is counting lattice points lying on or inside the ellipsoid $q(x_1, \ldots, x_n) = N$ in n-dimensional Euclidean space. Recalling our previous study of this sort of problem, we know that there exists a constant V such that

$$R_q(N) \sim V \cdot N^{n/2},$$

so that the average value of $r_q(N)$ is asymptotically $N^{\frac{n}{2}-1}$.

To say that $q(x_1, \ldots, x_n) = N$ has an integral solution is to say that $r_q(N) > 0$. It turns out to be a good strategy to exchange our problem for a seemingly harder problem: what can one say about the order of magnitude of $r_q(N)$?

One has the following theorem, thanks to the combined work of many leading mathematicians over a period of about 50 years:

THEOREM 19.10. *(Hecke, Eichler, Tartakowsky, Kloosterman, Deligne, ...) Suppose $q(x_1, \ldots, x_n)$ is positive definite and $n \geq 5$. There exists a decomposition*

$$r_q(N) = r_E(N) + r_C(N)$$

with the following properties:
a) $r_E(N) > 0$ iff the equation $q(x_1, \ldots, x_n) = N$ has solutions everywhere locally.
b) There exist effectively computable positive constants C_1, C_2 (depending on q) such that:

$$r_E(N) > 0 \implies r_E(N) \geq C_1 N^{n/2-1}.$$
$$|r_C(N)| \leq C_2 d(N) N^{\frac{n}{4}-\frac{1}{2}}.$$

Here $d(N)$ is the divisor function, which recall, grows slower than any positive power of N. One can interpret this result as saying that a local-global principle for $r_q(N)$ holds *asymptotically*, with **almost square root error**!

The proof of this theorem requires lots of techniques from 20th century number theory, and in particular the introduction of objects which are a lot less elementary and quaint than quadratic polynomials with integer coefficients. Notably the proof first associates to a quadratic form a **modular form** – a certain especially nice kind of function of a complex variable – and the result follows from a bound on the coefficients of a power series expansion of this function. In particular, one uses results on the number of solutions to much more general systems of equations over finite fields established by fundamental work of Pierre Deligne in the 1970's (work that justly landed him the Fields Medal).

COROLLARY 19.11. *Let q be a positive-definite quadratic form in $n \geq 5$ variables. Then there exists N_0 such that if $N \geq N_0$, $q(x_1, \ldots, x_n) = N$ satisfies the local-global principle (has integral solutions iff it has congruential solutions).*

Again, the theory of congruential solutions is sufficiently well-developed so as to enable one to determine (with some work, to be sure) precise conditions on N such that solutions exist everywhere locally. Therefore the corollary gives a method for solving the representation problem for integral quadratic forms in at least four variables: (i) explicitly compute the value of N_0 in the Corollary; (ii) explicitly compute the local conditions for solvability; (iii) check each of the finitely many values of N, $1 \leq N \leq N_0$ to see whether $q(x_1, \ldots, x_n) = N$ has a solution.

Thus the representation problem is reduced to a finite calculation. Of course not all finite problems can be solved in a reasonable (or even unreasonable) amount of time in practice, so quite a lot of technique and ingenuity is necessary to apply this method. Here is a success story:

THEOREM 19.12. *(Hanke* [**Han04**]*) The quadratic form* $x^3 + 3y^2 + 5z^2 + 7w^2$ *integrally represents all positive integers except* 2 *and* 22*.*

This result had been conjectured by M. Kneser in 1961.

Note that in Theorem 19.10 the number of variables has to be at least 4. When $n = 2$ or 3, the above corollary is false: we already mentioned this in the case of 2 variables, which is in some sense the hardest but also the best understood in terms of pure algebraic number theory. The case of ternary quadratic forms brings several new features and remains fascinatingly open. If you want to hear more, you will have to wait until 2008 and ask Prof. Hanke about it.

4. The 15 and 290 Theorems

The constants in Theorem 19.10 most definitely depend on the quadratic form q in question. A greater challenge is to prove results about integral representability that are in some sense independent of the particular quadratic form. For instance, a positive-definite quadratic form is said to be **universal** if it integrally represents every positive integer. (So the four squares form is universal.) The preceding section asserts the existence of a complicated procedure that can determine whether a given form is universal. Is there some easy way to determine whether a quadratic form is universal?

Yes. In the early 1990's, Conway and Schneeburger proved the following result.

THEOREM 19.13. *(15 Theorem* [**Con00**]*) A positive definite quadratic form with integral defining matrix integrally represents every positive integer iff it integrally represents the integers* 1 *through* 15*.*

Example: We will determine all positive integers d for which the form

$$x^2 + y^2 + z^2 + dw^2$$

is universal. We know that by taking $w = 0$ we can get every positive integer except those of the form $4^a(8k + 7)$; but since we need only go up to 15 it suffices to check whether we can represent 7. Let's check:

$d = 1$: $1^2 + 1^2 + 1^2 + 1 \cdot 2^2 = 7$.
$d = 2$: $2^2 + 1^2 + 0^2 + 2 \cdot 1^2 = 7$.
$d = 3$: $2^2 + 1^2 + 1^2 + 3 \cdot 1^2 = 7$.
$d = 4$: $1^2 + 1^2 + 1^2 + 4 \cdot 1^2 = 7$.
$d = 5$: $1^2 + 1^2 + 0^2 + 5 \cdot 1^2 = 7$.
$d = 6$: $1^2 + 0^2 + 0^2 + 6 \cdot 1^2 = 7$.
$d = 7$: $0^2 + 0^2 + 0^2 + 7 \cdot 1^2 = 7$.

We cannot represent 7 if $d \geq 8$: taking $w \neq 0$ would make the form too large.

In fact, let us consider the problem of which quadratic forms

$$q(x_1, x_2, x_3, x_4) = ax_1^2 + bx_2^2 + cx_3^2 + dx_4^2$$

with $a \leq b \leq c \leq d$ represent all positive integers. A case-by-case analysis shows that in order for the integers 1, 2, 3 and 5 to all be represented, we need (a, b, c) to be one of: $(1, 1, 1)$, $(1, 1, 2)$, $(1, 1, 3)$, $(1, 2, 2)$, $(1, 2, 3)$, $(1, 2, 4)$, $(1, 2, 5)$. As it happens, no ternary quadratic form can represent all positive integers. In the cases at hand, the smallest exceptions are (as you can readily check):

$x^2 + y^2 + z^2$ does not represent 7.
$x^2 + y^2 + 2z^2$ does not represent 14.
$x^2 + y^2 + 3z^2$ does not represent 6.
$x^2 + 2y^2 + 2z^2$ does not represent 7.
$x^2 + 2y^2 + 3z^2$ does not represent 10.
$x^2 + 2y^2 + 4z^2$ does not represent 14.
$x^2 + 2y^2 + 5z^2$ does not represent 10.

Now one can go through a similar analysis for the other 6 cases as we did for the first case, and determine a complete list of diagonal positive definite quaternary universal quadratic forms: there are precisely 54 of them.[3] In fact this investigation was originally done by S. Ramanujan in 1917, except that not having the 15 theorem he was forced to come up with "empirical" (i.e., conjectural) rules for which integers are represented by the above ternary quadratic forms, so that he did not supply proofs for his results.

Remark 4: Given the stories that have been told about Ramanujan and his unearthly intuition, it is interesting to remark that his paper lists a 55th universal quadratic form: $x^2 + 2y^2 + 5z^2 + 5w^2$. Ironically, this form does not represent 15, as Dickson observed ten years later.

The 15 theorem was discovered in a graduate seminar that Conway was teaching at Princeton, in which Schneeburger was an attending student. The original proof was quite computationally onerous, and it was never written down. Indeed, by the time Manjul Bhargava became a graduate student at Princeton and heard about the theorem, some of the details of the proof had been forgotten.

Bhargava was doubly stunned by this: that such a wonderful theorem could have been discovered, and also that it had met such a disappointing fate. He found a new proof of the 15 theorem which is, truly, one of the most beautiful mathematical arguments I have ever seen. It quite cleverly manages to avoid any unwieldy computations. In fact he proved the following generalization:

THEOREM 19.14. *(Bhargava's Master Theorem) Let $S \subset \mathbb{Z}^+$. There exists a finite subset S_0 of S such that a positive definite integer-matrix quadratic form represents all integers in S iff it represents all integers in S_0.*

[3]It can now be told that I put this as an extra credit problem on the final exam. Moreover, I hinted that I might do so, and in fact there was a student who practiced this type of calculation and was able to give the complete solution!

Example: Taking S to be the prime numbers, Bhargava showed that one may take S_0 to be the primes less than or equal to 73.

The proof gives an algorithm for determining S_0, but whether or not it is practical seems to depend very much on the choice of S: it gets much harder if S does not contain several very small integers.

Indeed, we have been saying "integer matrix" quadratic forms for the last few results, but a quadratic form is represented by a polynomial with integer coefficients iff its defining matrix satisfies the slightly weaker condition that its diagonal entries are integers and its off-diagonal entries are half-integers (e.g. $q(x, y) = xy$). However, if q is any integral quadratic form, then the matrix entries of $2q$ are certainly integers, and q represents an integer N iff $2q$ represents $2N$. Thus, applying Bhargava's Master Theorem to the subset of positive *even* integers, one deduces the existence of an integer N_0 such that if a positive-definite integral matrix represents every $N \in \{1, \ldots, N_0\}$ then it represents every positive integer.

Already in Conway's course it was suggested that N_0 could be taken to be 290. However, the calculations necessary to establish this result were Herculean: one needs to show that each of $6,436$ quaternary quadratic forms is universal. Some of these forms can be proven universal in relatively slick and easy ways, but about $1,000$ of them are seriously hard. So Bhargava enlisted the help of Jonathan Hanke, and after several years of intense work (including extremely intensive and carefully checked computer calculations), they were able to show the following result.

THEOREM 19.15. *(290 Theorem* [**BHxx**]*) If a positive-definite integral quadratic form represents each of:*

$$1, 2, 3, 5, 6, 7, 10, 13, 14, 15, 17, 19, 21, 22, 23, 26, 29, 30, 31, 34, 35, 37, 42, 58, 93, 110, 145, 203, 290,$$

then it represents all positive integers.

APPENDIX A

Rings, Fields and Groups

1. Rings

Recall that a binary operation on a set S is just a function $* : S \times S \to S$: in other words, given any two elements $s_1,\ s_2$ of S, there is a well-defined element $s_1 * s_2$ of S.

A **ring** is a set R endowed with two binary operations $+$ and $\cdot$, called addition and multiplication, respectively, which are required to satisfy a rather long list of familiar-looking conditions – in all the conditions below, a, b, c denote arbitrary elements of R –

(A1) $a + b = b + a$ (commutativity of addition);
(A2) $(a + b) + c = a + (b + c)$ (associativity of addition);
(A3) There exists an element, called 0, such that $0 + a = a$. (additive identity)
(A4) For $x \in R$, there is a $y \in R$ such that $x + y = 0$ (existence of additive inverses).
(M1) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ (associativity of multiplication).
(M2) There exists an element, called 1, such that $1 \cdot a = a \cdot 1 = a$.
(D) $a \cdot (b + c) = a \cdot b + a \cdot c$; $(a + b) \cdot c = a \cdot c + b \cdot c$.

Comments:

(i) The additive inverse required to exist in (A4) is unique, and the additive inverse of a is typically denoted $-a$. (It is easy to check that $-a = (-1) \cdot a$.)

(ii) Note that we require the existence of a multiplicative identity (or a "unity"). Every once in a while one meets a structure which satisfies all the axioms except does not have a multiplicative identity, and one does not eject it from the club just because of this. But all of our rings will have a multiplicative identity.

(iii) There are two further reasonable axioms on the multiplication operation that we have *not* required; our rings will sometimes satisfy them and sometimes not:

(M′) $a \cdot b = b \cdot a$ (commutativity of multiplication).
(M″) For all $a \neq 0$, there exists $b \in R$ such that $ab = 1$.

A ring which satisfies (M′) is called – sensibly enough – a **commutative ring**.

Example 1.0: The integers $\mathbb{Z}$ form a ring under addition and multiplication. Indeed they are "the universal ring" in a sense to be made precise later.

Example 1.1: There is a unique ring in which $1 = 0$. Indeed, if r is any element of such a ring, then

$$r = 1 \cdot r = 0 \cdot r = (0 + 0) \cdot r = 0 \cdot r + 0 \cdot r = 1 \cdot r + 1 \cdot r = r + r;$$

subtracting r from both sides, we get $r = 0$. In other words, the only element of the ring is 0 and the addition laws are just $0 + 0 = 0 = 0 \cdot 0$; this satisfies all the axioms for a commutative ring. We call this the **zero ring**. Truth be told, it is a bit of annoyance: often in statements of theorems one encounters "except for the zero ring."

Example 1.n: For any positive integer, let Z_n denote the set $\{0, 1, \ldots, n-1\}$. There is a function mod_n from the positive integers to Z_n: given any integer m, $\mathrm{mod}_n(m)$ returns the remainder of m upon division by n, i.e., the unique integer r satisfying $m = qn + r$, $0 \leq r < n$. We then define operations of $+$ and $\cdot$ on Z_n by viewing it as a subset of the positive integers, employing the standard operations of $+$ and $\cdot$, and then applying the function mod_n to force the answer back in the range $0 \leq r < n$. That is, we define

$$a +_n b := \mathrm{mod}_n(a + b),$$

$$a \cdot_n b := \mathrm{mod}_n(a \cdot b).$$

The addition operation is familiar from "clock arithmetic": with $n = 12$ this is how we tell time, except that we use $1, 2, \ldots, 12$ instead of $0, \ldots, 11$. (However, military time does indeed go from 0 to 23.)

The (commutative!) rings Z_n are basic and important in all of mathematics, especially number theory. The definition we have given – the most "naive" possible one – is not quite satisfactory: how do we know that $+_n$ and $\cdot_n$ satisfy the axioms for a ring? Intuitively, we want to say that the integers $\mathbb{Z}$ form a ring, and the Z_n's are constructed from $\mathbb{Z}$ in some way so that the ring axioms become automatic. This leads us to the *quotient construction*, which we will present later.

Modern mathematics has tended to explore the theory of commutative rings much more deeply and systematically than the theory of (arbitrary) non-commutative rings. Nevertheless noncommutative rings are important and fundamental: the basic example is the ring of $n \times n$ matrices (say, with real entries) for any $n \geq 2$.

A ring (except the zero ring!) which satisfies (M″) is called a **division ring** (or division algebra). Best of all is a ring which satisfies (M′) and (M″): a **field**.[1]

I hope you have some passing familiarity with the fields $\mathbb{Q}$ (of rational numbers), $\mathbb{R}$ (of real numbers) and $\mathbb{C}$ (of complex numbers), and perhaps also with the existence of finite fields of prime order (more on these later). In some sense a field is the richest possible purely algebraic structure, and it is tempting to think of the elements

[1] A very long time ago, some people used the term "field" as a synonym for "division ring" and therefore spoke of "commutative fields" when necessary. The analogous practice in French took longer to die out, and in relatively recent literature it was not standardized whether "corps" meant any division ring or a commutative division ring. (One has to keep this in mind when reading certain books written by Francophone authors and less-than-carefully translated into English, e.g. Serre's *Corps Locaux*.) However, the Bourbakistic linguistic philosophy that the more widely used terminology should get the simpler name seems to have at last persuaded the French that "corps" means "(commutative!) field."

of field as "numbers" in some suitably generalized sense. Conversely, elements of arbitrary rings can have some strange properties that we would, at least initially, not want "numbers" to have.

2. Ring Homomorphisms

Generally speaking, a **homomorphism** between two algebraic objects is a map f between the underlying sets which preserves all the relevant algebraic structure.

So a **ring homomorphism** $f : R \to S$ is a map such that $f(0) = 0$, $f(1) = 1$ and for all $r_1,\ r_2 \in R$, $f(r_1 + r_2) = f(r_1) + f(r_2)$, $f(r_1 r_2) = f(r_1)f(r_2)$.

In fact it follows from the preservation of addition that $f(0) = 0$. Indeed, $0 = 0+0$, so $f(0) = f(0+0) = f(0) + f(0)$; now subtract $f(0)$ from both sides. But in general it seems better to postulate that a homomorphism preserve every structure "in sight" and then worry later about whether any of the preservation properties are redundant. Note well that the property $f(1) = 1$ – "unitality" – is *not* redundant. Otherwise every ring R would admit a homomorphism to the zero ring, which would turn out to be a bit of a pain.

Example 2.1: For any ring R, there exists a unique homomorphism $c : \mathbb{Z} \to R$. Namely, any homomorphism must send 1 to 1_R, 2 to $1_R + 1_R$, 3 to $1_R + 1_R + 1_R$, -1 to -1_R, -2 to $-1_R + -1_R$ and so forth. (And it is not hard to see that this necessarily gives a homomorphism.)

Recall that a function $f : X \to Y$ is an **injection** if $x_1 \neq x_2 \implies f(x_1) \neq f(x_2)$. To see whether a homomorphism of rings $f : R \to S$ is an injection, it suffices to look at the set $K(f) = \{x \in R \mid f(x) = 0\}$, the **kernel** of f. This set contains 0, and if it contains any other element then f is certainly not injective. The converse is also true: suppose $K(f) = 0$ and $f(x_1) = f(x_2)$. Then $0 = f(x_2) - f(x_1) = f(x_2 - x_1)$, so $x_2 - x_1 \in K(f)$, so by our assumption $x_2 - x_1 = 0$, and $x_1 = x_2$.

An important case is when R is a ring and S is a subset of R containing 0 and 1 and which is itself a ring under the operations of $+$ and $\cdot$ it inherits from R. (In this case what needs to be checked are the *closure* of S under $+$, $-$ and $\cdot$: i.e., for all $s_1,\ s_2 \in S$, $s_1 + s_2,\ s_1 - s_2,\ s_1 \cdot s_2 \in S$.) We say that S is a **subring** of R.

Suppose R and S are division rings and $f : R \to S$ is a homomorphism between them. Suppose that r is in the kernel of f, i.e., $f(r) = 0$. If $r \neq 0$, then it has a (left and right) multiplicative inverse, denoted r^{-1}, i.e., an element such that $rr^{-1} = r^{-1}r = 1$. But then

$$1 = f(1) = f(rr^{-1}) = f(r)f(r^{-1}) = 0 \cdot f(r^{-1}) = 0,$$

a contradiction. So any homomorphism of division rings is an injection: it is especially common to speak of **field extensions**. For example, the natural inclusions $\mathbb{Q} \hookrightarrow \mathbb{R}$ and $\mathbb{R} \hookrightarrow \mathbb{C}$ are both field extensions.

Example 2.1, continued: recall we have a unique homomorphism $c : \mathbb{Z} \to R$. If c is injective, then we find a copy of the integers naturally as a subring of R. E.g. this is the case when $R = \mathbb{Q}$. If not, there exists a least positive integer n such that

$c(n) = 0$, and one can check that $\mathrm{Ker}(c)$ consists of all integer multiples of n, a set which we will denote by $n\mathbb{Z}$ or by (n). This integer n is called the **characteristic** of R, and if no such n exists we say that R is of characteristic 0 (yes, it would seem to make more sense to say that n has infinite characteristic). As an important example, the homomorphism $c : \mathbb{Z} \to Z_n$ is an extension of the map mod_n to all of $\mathbb{Z}$; in particular the characteristic of Z_n is n.

3. Integral Domains

A commutative ring R (which is not the zero ring!) is said to be an **integral domain** if it satisfies either of the following *equivalent* properties:[2]

(ID1) If $x, y \in R$ and $xy = 0$ then $x = 0$ or $y = 0$.
(ID2) If $a, b, c \in R$, $ab = ac$ and $a \neq 0$, then $b = c$.

(Suppose R satisfies (ID1) and $ab = ac$ with $a \neq 0$. Then $a(b - c) = 0$, so $b - c = 0$ and $b = c$; so R satisfies (ID2). The converse is similar.)

(ID2) is often called the "cancellation" property and it is extremely useful when solving equations. Indeed, when dealing with equations in a ring which is not an integral domain, one must remember *not* to apply cancellation without further justification! (ID1) expresses the nonexistence of **zero divisors**: a nonzero element x of a ring R is called a zero divisor if there exists y in R such that $xy = 0$.

An especially distressing kind of zero divisor is an element $0 \neq a \in R$ such that $a^n = 0$ for some positive integer n. (If N is the least positive integer N such that $a^N = 0$ we have $a,\ a^{N-1} \neq 0$ and $a \cdot a^{N-1} = 0$, so a is a zero divisor.) Such an element is called **nilpotent**, and a ring is **reduced** if it has no nilpotent elements.

One of the difficulties in learning ring theory is that the examples have to run very fast to keep up with all the definitions and implications among definitions. But, look, here come some now:

Example 3.1: Let us consider the rings Z_n for the first few n.

The rings Z_2 and Z_3 are easily seen to be fields: indeed, in Z_2 the only nonzero element, 1 is its own multiplicative inverse, and in Z_3 $1 = 1^{-1}$ and $2 = (2)^{-1}$.

In the ring Z_4 $2^2 = 0$, so 2 is nilpotent and Z_4 is nonreduced.

In Z_5 one finds – after some trial and error – that $1^{-1} = 1,\ 2^{-1} = 3,\ 3^{-1} = 2,\ 4^{-1} = 4$ so that Z_5 is a field.

In Z_6 we have $2 \cdot 3 = 0$ so there are zero-divisors, but a bit of calculation shows there are no nilpotent elements. (We take enough powers of every element until we get the same element twice; if we never get zero then no power of that element will be zero. For instance $2^1 = 2$, $2^2 = 4$, $2^3 = 2$, so 2^n will equal either 2 or 4 in Z_6: never 0.)

[2]The terminology "integral domain" is completely standardized but a bit awkward: on the one hand, the term "domain" has no meaning by itself. On the other hand there is also a notion of an "integral extension of rings." And, alas, it may well be the case that an extension of integral domains is not an integral extension! But there is no clear remedy here, and proposed changes in the terminology – e.g. Lang's attempted use of "entire" for "integral domain" – have not been well received.

Similarly we find that Z_7 is a field, 2 is a nilpotent in Z_8, 3 is a nilpotent in Z_9, Z_{10} is reduced but not an integral domain, and so forth. Eventually it will strike us that it appears to be the case that Z_n is a field exactly when n is prime. This realization makes us pay closer attention to the prime factorization of n, and given this clue, one soon guesses that Z_n is reduced iff n is **squarefree**, i.e., not divisible by the square of any prime. Moreover, it seems that whenever Z_n is an integral domain, it is also a field. All of these observations are true in general but nontrivial to prove. The last fact is the easiest:

PROPOSITION A.1. *Any integral domain R with finitely many elements is a field.*

PROOF. Consider any $0 \neq a \in R$; we want to find a multiplicative inverse. Consider the various powers $a^1,\ a^2, \ldots$ of a. They are, obviously, elements of R, and since R is finite we must eventually get the same element via distinct powers: there exist positive integers i and j such that $a^{i+j} = a^i \neq 0$. But then $a^i = a^{i+j} = a^i \cdot a^j$, and applying (ID2) we get $a^j = 1$, so that a^{j-1} is the multiplicative inverse to a. □

THEOREM A.2. *a) The ring Z_n is a field iff n is a prime.*
b) The ring Z_n is reduced iff n is squarefree.

PROOF. In each case one direction is rather easy. Namely, if n is not prime, then $n = ab$ for integers $1 < a,\ b < n$, and then $a \cdot b = 0$ in Z_n. If n is not squarefree, then for some prime p we can write $n = p^2 \cdot m$, and then the element mp is nilpotent: $(mp)^2 = mp^2m = mn = 0$ in Z_n.

However, in both cases the other direction requires Euclid's Lemma: if a prime p divides ab then $p|a$ or $p|b$. (We will encounter and prove this result early on in the course.) Indeed, this says precisely that if $ab = 0$ in Z_n then either $a = 0$ or $b = 0$, so Z_p is an integral domain, and, being finite, by Proposition A.1 it is then necessarily a field. Finally, if $n = p_1 \cdots p_n$ is squarefree, and $m < n$, then m is not divisible by some prime divisor of n, say p_i, and by the Euclid Lemma neither is any power m^a of m, so for no positive integer a is $m^a = 0$ in Z_n. □

Example 3.2: Of course, the integers $\mathbb{Z}$ form an integral domain. How do we know? Well, if $a \neq 0$ and $ab = ac$, we can multiply both sides by a^{-1} to get $b = c$. This may at first seem like cheating since a^{-1} is generally not an integer: however it exists as a rational number and the "computation" makes perfect sense in $\mathbb{Q}$. Since $\mathbb{Z} \subset \mathbb{Q}$, having $b = c$ in $\mathbb{Q}$ means that $b = c$ in $\mathbb{Z}$.

It turns out that for *any* commutative ring R, if R is an integral domain we can prove it by the above argument of exhibiting a field that contains it:

THEOREM A.3. *A ring R is an integral domain iff it is a subring of some field.*

PROOF. The above argument (i.e., just multiply by a^{-1} in the ambient field) shows that any subring of a field is an integral domain. The converse uses the observation that given an integral domain R, one can formally build a field $F(R)$ whose elements are represented by formal fractions of the form $\frac{a}{b}$ with $a \in R,\ b \in R \setminus \{0\}$, subject to the rule that $\frac{a}{b} = \frac{c}{d}$ iff $ad = bc$ in R. There are many little checks to make to see that this construction actually works. On the other hand, this is a direct generalization of the construction of the field $\mathbb{Q}$ from the integral domain $\mathbb{Z}$, so we feel relatively sanguine about omitting the details here. □

Remark: The field $F(R)$ is called the **field of fractions**[3] of the integral domain R.

Example 3.3 (Subrings of $\mathbb{Q}$): There are in general many different integral domains with a given quotient field. For instance, let us consider the integral domains with quotient field $\mathbb{Q}$, i.e., the subrings of $\mathbb{Q}$. The two obvious ones are $\mathbb{Z}$ and $\mathbb{Q}$, and it is easy to see that they are the extremes: i.e., for any subring R of $\mathbb{Q}$ we have $\mathbb{Z} \subseteq R \subseteq \mathbb{Q}$. But there are many others: for instance, let p be any prime number, and consider the subset R_p of $\mathbb{Q}$ consisting of rational numbers of the form $\frac{a}{b}$ where b is not divisible by any prime except p (so, taking the convention that $b > 0$, we are saying that $b = p^k$ for some k). A little checking reveals that R_p is a subring of $\mathbb{Q}$. In fact, this construction can be vastly generalized: let S be any subset of the prime numbers (possibly infinite!), and let R_S be the rational numbers $\frac{a}{b}$ such that b is divisible only by primes in S. It is not too hard to check that: (i) R_S is a subring of $\mathbb{Q}$, (ii) if $S \neq S'$, $R_S \neq R_{S'}$, and (iii) every subring of $\mathbb{Q}$ is of the form R_S for some set of primes S. Thus there are uncountably many subrings in all!

4. Polynomial Rings

Let R be a commutative ring. One can consider the ring $R[T]$ of polynomials with coefficients in T: that is, the union over all natural numbers n of the set of all formal expressions $\sum_{i=0}^{n} a_i T^i$ ($T^0 = 1$). (If $a_n \neq 0$, this polynomial is said to have degree n. By convention, we take the zero polynomial to have degree $-\infty$.) There are natural addition and multiplication laws which reduce to addition in R, the law $T^i \cdot T^j = T^{i+j}$ and distributivity. (Formally speaking we should write down these laws precisely and verify the axioms, but this is not very enlightening.) One gets a commutative ring $R[T]$.

One can also consider polynomial rings in more than one variable: $R[T_1, \ldots, T_n]$. These are what they sound like; among various possible formal definitions, the most technically convenient is an inductive one: $R[T_1, \ldots, T_n] := R[T_1, \ldots, T_{n-1}][T_n]$, so e.g. the polynomial ring $R[X, Y]$ is just a polynomial ring in one variable (called Y) over the polynomial ring $R[X]$.

PROPOSITION A.4. *$R[T]$ is an integral domain iff R is an integral domain.*

PROOF. R is naturally a subring of $R[T]$ – the polynomials rT^0 for $r \in R$ and any subring of an integral domain is a domain; this shows necessity. Conversely, suppose R is an integral domain; then any two nonzero polynomials have the form $a_n T^n + a_{n-1} T^{n-1} + \ldots + a_0$ and $b_m T^m + \ldots + b_0$ with $a_n,\ b_m \neq 0$. When we multiply these two polynomials, the leading term is $a_n b_m T^{n+m}$; since R is a domain, $a_n b_m \neq 0$, so the product polynomial has nonzero leading term and is therefore nonzero. □

COROLLARY A.5. *A polynomial ring in any number of variables over an integral domain is an integral domain.*

This construction gives us many "new" integral domains and hence many new fields. For instance, starting with a field F, the fraction field of $F[T]$ is the set of all formal

[3]The term "quotient field" is also used, even by me until rather recently. But since there is already a quotient construction in ring theory, it seems best to use a different term for the fraction construction.

quotients $\frac{P(T)}{Q(T)}$ of polynomials; this is denoted $F(T)$ and called the field of rational functions over F. (One can equally well consider fields of rational functions in several variables, but we shall not do so here.)

The polynomial ring $F[T]$, where F is a field, has many nice properties; in some ways it is strongly reminiscent of the ring $\mathbb{Z}$ of integers. The most important common property is the ability to divide:

THEOREM A.6. *(Division theorem for polynomials) Given any two polynomials $a(T)$, $b(T)$ in $F[T]$, there exist unique polynomials $q(T)$ and $r(T)$ such that*

$$b(T) = q(T)a(T) + r(T)$$

and $\deg(r(T)) < \deg(a(T))$.

A more concrete form of this result should be familiar from high school algebra: instead of formally proving that such polynomials exist, one learns an algorithm for actually finding $q(T)$ and $r(T)$. Of course this is as good or better: all one needs to do is to give a rigorous proof that the algorithm works, a task we leave to the reader. (Hint: induct on the degree of b.)

COROLLARY A.7. *(Factor theorem) For $a(T) \in F[T]$ and $c \in F$, the following are equivalent:*
a) $a(c) = 0$.
b) $a(T) = q(T) \cdot (T - c)$.

PROOF. We apply the division theorem with $b(T) = (T - c)$, getting $a(T) = q(T)(T - c) + r(T)$. The degree of r must be less than the degree of $T - c$, i.e., zero – so r is a constant. Now plug in $T = c$: we get that $a(c) = r$. So if $a(c) = 0$, $a(T) = q(T)(T - c)$. The converse is obvious. □

COROLLARY A.8. *A nonzero polynomial $p(T) \in F[T]$ has at most* $\deg(p(T))$ *roots.*

Remark: The same result holds for polynomials with coefficients in an integral domain R, since every root of p in R is also a root of p in the fraction field $F(R)$.

This may sound innocuous, but do not underestimate its power – a judicious application of this Remark (often in the case $R = \mathbb{Z}/p\mathbb{Z}$) can and will lead to substantial simplifications of "classical" arguments in elementary number theory.

Example 4.1: Corollary A.8 does *not* hold for polynomials with coefficients in an arbitrary commutative ring: for instance, the polynomial $T^2 - 1 \in Z_8[T]$ has degree 2 and 4 roots: $1, 3, 5, 7$.

5. Commutative Groups

A **group** is a set G endowed with a single binary operation $* : G \times G \to G$, required to satisfy the following axioms:

(G1) for all $a, b, c \in G$, $(a * b) * c = a * (b * c)$ (associativity)
(G2) There exists $e \in G$ such that for all $a \in G$, $e * a = a * e = a$.
(G3) For all $a \in G$, there exists $b \in G$ such that $ab = ba = e$.

Example: Take an arbitrary set S and put $G = \mathrm{Sym}(S)$, the set of all bijections $f : S \to S$. When $S = \{1, \ldots, n\}$, this is called the **symmetric group** of order n, otherwise known as the group of all permutations on n elements: it has order $n!$.

We have notions of **subgroups** and **group homomorphisms** that are completely analogous to the corresponding ones for rings: a subgroup $H \subset G$ is a subset which is nonempty, and is closed under the group law and inversion: i.e., if $g,\ h \in H$ then also $g * h$ and g^{-1} are in H. (Since there exists some $h \in H$, also h^{-1} and $e = h * h^{-1} \in H$; so subgroups necessarily contain the identity.)[4] And a homomorphism $f : G_1 \to G_2$ is a map of groups which satisfies $f(g_1 * g_2) = f(g_1) * f(g_2)$ (as mentioned above, that $f(e_{G_1}) = e_{G_2}$ is then automatic). Again we get many examples just by taking a homomorphism of rings and forgetting about multiplication.

Example 5.1: Let F be a field. Recall that for any positive integer n, the $n \times n$ matrices with coefficients in F form a ring under the operations of matrix addition and matrix multiplication, denoted $M_n(F)$. Consider the subset of invertible matrices, $GL_n(F)$. It is easy to check that the invertible matrices form a group under matrix multiplication (the "unit group" of the ring $M_n(F)$, coming up soon). No matter what F is, this is an interesting and important group, and is not commutative if $n \geq 2$ (when $n = 1$ it is just the group of nonzero elements of F under multiplication). The determinant is a map

$$\det : GL_n(F) \to F \setminus \{0\};$$

a well-known property of the determinant is that $\det(AB) = \det(A)\det(B)$. In other words, the determinant is a homomorphism of groups. Moreover, just as for a homomorphism of rings, for any group homomorphism $f : G_1 \to G_2$ we can consider the subset $K_f = \{g \in G_1 \mid f(g) = e_{G_2}\}$ of elements mapping to the identity element of G_2, again called the **kernel** of f. It is easy to check that K_f is always a subgroup of G_1, and that f is injective iff $K_f = 1$. The kernel of the determinant map is denoted $SL_n(F)$; by definition, it is the collection of all $n \times n$ matrices with determinant 1.[5] For instance, the rotation matrices $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ form a subset (indeed, a subgroup) of the group $SL_2(\mathbb{R})$.

THEOREM A.9. *(Lagrange) For a subgroup H of the finite group G, we have $\#H \mid \#G$.*

The proof[6] is combinatorial: we exhibit a partition of G into a union of subsets H_i, such that $\#H_i = \#H$ for all i. Then, the order of G is $\#H \cdot n$, where n is the number of subsets.

The H_i's will be the **left cosets** of H, namely the subsets of the form

$$gH = \{gh \mid h \in H\}.$$

[4]Indeed there is something called the "one step subgroup test": a nonempty subset $H \subset G$ is a subgroup iff whenever g and h are in H, then $g * h^{-1} \in H$. But this is a bit like saying you can put on your pants in "one step" if you hold them steady and jump into them: it's true but not really much of a time saver.

[5]The "GL" stands for "general linear" and the "SL" stands for "special linear."

[6]This proof may be too brief if you have not seen the material before; feel free to look in any algebra text for more detail, or just accept the result on faith for now.

Here g ranges over all elements of G; the key is that for $g_1,\ g_2 \in G$, the two cosets g_1H and g_2H are either equal or disjoint – i.e., what is not possible is for them to share some but not all elements. To see this: suppose $x \in g_1H$ and is also in g_2H. This means that there exist $h_1,\ h_2 \in H$ such that $x = g_1h_1$ and also $x = g_2h_2$, so $g_1h_1 = g_2h_2$. But then $g_2 = g_1h_1h_2^{-1}$, and since $h_1,\ h_2 \in H$, $h_3 := h_1h_2^{-1}$ is also an element of H, meaning that $g_2 = g_1h_3$ is in the coset g_1H. Moreover, for any $h \in H$, this implies that $g_2h = g_1h_3h = g_1h_4 \in g_1H$, so that $g_2H \subset g_1H$. Interchanging the roles of g_2 and g_1, we can equally well show that $g_1H \subset g_2H$, so that $g_1H = g_2H$. Thus overlapping cosets are equal, which was to be shown.

Remark: In the proof that G is partitioned into cosets of H, we did not use the finiteness anywhere; this is true for all groups. Indeed, for any subgroup H of any group G, we showed that there is a set S – namely the set of distinct left cosets $\{gH\}$ such that the elements of G can be put in bijection with $S \times H$. If you know about such things (no matter if you don't), this means precisely that $\#H$ divides $\#G$ even if one or more of these cardinalities is *infinite*.

COROLLARY A.10. *If G has order n, and $g \in G$, then the order of g – i.e., the least positive integer k such that $g^k = 1$ – divides n.*

PROOF. The set of all positive powers of an element of a finite group forms a subgroup, denoted $\langle g \rangle$, and it is easily checked that the distinct elements of this group are $1,\ g,\ g^2,\ \ldots, g^{k-1}$, so the order of g is also $\#\langle g \rangle$. Thus the order of g divides the order of G by Lagrange's Theorem. □

Example 5.2: For any ring R, $(R, +)$ is a commutative group. Indeed, there is nothing to check: a ring is simply more structure than a group. For instance, we get for each n a commutative group Z_n just by taking the ring Z_n and forgetting about the multiplicative structure.

A group G is called **cyclic** if it has an element g such that every element x in G is of the form $1 = g^0$, $g^n := g \cdot g \cdot \ldots \cdot g$ or $g^{-n} = g^{-1} \cdot g^{-1} \cdot \ldots \cdot g^{-1}$ for some positive integer n. The group $(\mathbb{Z}, +)$ forms an infinite cyclic group; for every positive integer n, the group $(Z_n, +)$ is cyclic of order n. It is not hard to show that these are the only cyclic groups, up to isomorphism.

An element u of a ring R is a **unit** if there exists $v \in R$ such that $uv = vu = 1$.

Example 5.3: 1 is always a unit; 0 is never a unit (except in the zero ring, in which $0 = 1$). The units in $\mathbb{Z}$ are ± 1.

A nonzero ring is a division ring iff every nonzero element is a unit.

The set of all units in a ring is denoted $R^\times$. It is not hard to see that the units form a group under multiplication: for instance, if u and v are units, then they have two-sided inverses denoted u^{-1} and v^{-1}, and then

$$uv \cdot (v^{-1}u^{-1}) = (v^{-1}u^{-1})uv = 1,$$

so uv is also a unit. Similarly, the (unique) inverse u^{-1} of a unit is a unit. In general, $R^\times$ is not commutative, but of course it will be if R is commutative.

6. Ideals and Quotients

In this section all groups and rings will be commutative.

Let R be a (commutative!) ring. An **ideal** of R is a subset I of R satisfying:
(IR1) I is a subgroup of the additive group of R.
(IR2) For any $r \in R$ and any $i \in I$, $ri \in I$.

We often employ notation like $rI = \{ri \mid i \in I\}$ and then (IR2) can be stated more succinctly as: for all $r \in R$, $rI \subset I$. In other words, an ideal is a subset of a ring R which is a subgroup under addition (in particular it contains 0 so is nonempty) and is not only closed under multiplication but satisfies the *stronger* property that it "absorbs" all elements of the ring under multiplication.

Remark (Ideals versus subrings): It is worthwhile to compare these two notions; they are related, but with subtle and important differences. Both an ideal I and a subring S of a ring R are subsets of R which are subgroups under addition and are stable under multiplication. However, each has an additional property: for an ideal it is the *absorption* property (IR2). For instance, the integers $\mathbb{Z}$ are a subring of the rational numbers $\mathbb{Q}$, but are clearly not an ideal, since $\frac{1}{2} \cdot 1 = \frac{1}{2}$, which is not an integer. On the other hand a subring has a property that an ideal usually lacks, namely it must contain the unity 1 of R. For instance, the subset $2\mathbb{Z} = \{2n \mid n \in \mathbb{Z}\}$ is an ideal of $\mathbb{Z}$ but is not a subring.

Example (trivial ideals): Any ring R (which is not the zero ring!) contains at least two ideals: the ideal $\{0\}$, and the ideal R itself. These are however not very interesting examples, and often need to be ignored in a discussion. (The convention that "ideal" should stand for "non-zero ideal" whenever convenient is a fairly common and useful one in the subject.) An ideal I is said to be **proper** if it is not R, and again most interesting statements about ideals should really be applied to proper ideals. Note well that an ideal is proper iff it does not contain the unity 1. Indeed, an ideal lacking 1 is certainly proper, and conversely, if $1 \in I$ and $r \in R$, then $r \cdot 1 = r$ is in R.

PROPOSITION A.11. *The following are equivalent for a nonzero commutative ring R:*
a) R has only the trivial ideals $\{0\}$ and R.
b) R is a field.

PROOF. b) $\implies$ a): Suppose I is a nonzero ideal of a field R, so I contains some $0 \neq a$. Then since a is a field, a^{-1} exists and $1 = a^{-1}a \in R \cdot I \subset I$, so I contains 1 and is hence all of R.

a) $\implies$ b): Suppose R is not a field; then some nonzero element a does not have an inverse. Then the set $aR = \{ar \mid r \in R\}$ is a proper, nonzero ideal. $\square$

The preceding argument shows a general construction of ideals: for any element a of R, the set of elements $\{ra \mid r \in R\}$ is an ideal of R. (Just to check: $r_1a + r_2a = (r_1 + r_2)a$, $-ra = (-r)a$ and $r'(ra) = (r'r)a$.) We denote such ideals as either Ra

or (a); they are simple and easy to understand and are called **principal ideals** and a is called a *generator*.

PROPOSITION A.12. *(To contain is to divide) Let a and b be elements of R. The following are equivalent:*
a) $(a) \supset (b)$.
b) $a \mid b$*; i.e.,* $b = ac$ *for some* c *in* R.

PROOF. Exercise! □

PROPOSITION A.13. *Let a and b be elements of an integral domain R. The following are equivalent:*
a) $(a) = (b)$.
b) $a = bu$ *for some unit* $u \in R^\times$.

PROOF. Since $(a) \supset (b)$, $b = c_1 a$ for some $c_1 \in R$. Since $(b) \supset (a)$, $a = c_2 b$ for some $c_2 \in R$. Combining this information we get $b = c_1 c_2 b$. If $b = 0$, then also $a = 0$, a trivial case; otherwise, we can cancel b to get $c_1 c_2 = 1$, meaning that c_1 and c_2 are units, so $a = c_2 b$ shows what we want. □

The notion of generators for ideals can be generalized: for any $a_1, \ldots, a_n$, there is an ideal, denoted $(a_1, \ldots, a_n)$, and defined as the set of all elements of the form $a_1 r_1 + \ldots + a_n r_n$ as $r_1, \ldots, r_n$ range over all elements of R. This is called the ideal *generated by* $a_1, \ldots, a_n$, and it is not hard to see that any ideal I which contains the elements $a_1, \ldots, a_n$ must contain the ideal $(a_1, \ldots, a_n)$.

Example: In particular, taking $R = \mathbb{Z}$, for any $n \in \mathbb{Z}$ we have an ideal (n) consisting of all integer multiples of n. Taking $n = 0$ we get the zero ideal; otherwise $(n) = (-n)$, so that the same principal ideal may well have more than one generator. (However, in this case we are quite fortunate and every nonzero principal ideal has a "standard" generator, namely the positive one.) Consider the ideal generated by 4 and 7; is there some simpler way of writing it? Indeed yes: since $1 = 3 \cdot 7 + (-5) \cdot 4$, the ideal $(4, 7)$ contains 1, so it is in fact all of $\mathbb{Z}$: $(4, 7) = (1) = \mathbb{Z}$. If you play around a bit with ideals generated by two different integers, you will soon come around to the idea that (a, b) is always principal, and is generated by the greatest common divisor of a and b. (Note that we are all of a sudden doing number theory!) However, this turns out to be quite nontrivial to prove; it is the essential content of the fundamental theorem of arithmetic.

Example: Let $R = \mathbb{Z}[T]$, and consider the ideal $(2, T)$: it consists of all polynomials of the form $2P(T) + TQ(T)$. Is this ideal principal? If so, there would exist a single magic polynomial $M(T)$ such that $2 = p_1(T) * M(T)$ and $T = p_2(T) * M(T)$. But since the degree of the product polynomial is the sum of the degrees of the two polynomials, we get that $p_1(T)$ and $M(T)$ are both constants, i.e., both integers. Replacing $M(T)$ by $-M(T)$ if necessary, we see that the only possibilities are $M(T) = 1$ or $M(T) = 2$. But if $M(T) = 2$, then $T = 2p_2(T)$ which is impossible, because 2 divides the leading term of $2p_2(T)$. So the only possibility is that $M(T) = 1$, i.e., perhaps the ideal $(2, T)$ is all of R? No, this isn't right either: it is not possible that $1 = 2P(T) + TQ(T)$: plugging in $T = 0$ we get that $1 = 2P(0)$, and again, 1 is not divisible by 2. The ideal $(2, T)$ is not principal – too bad!

Definition: A ring R is called *principal* if every ideal is principal. This property interacts especially nicely with integral domains: a domain in which each ideal is principal is called a **principal ideal domain**, or PID. Such are the rings whose arithmetic most closely parallels the arithmetic of the integers, so are especially prized.

6.1. Prime and maximal ideals. A proper ideal I in a ring is **prime** if whenever $xy \in I$, $x \in I$ or $y \in I$. A proper ideal is **maximal** if it is not strictly contained in any larger proper ideal.

The first definition requires some justification. At the moment, we ask the reader to consider which of the ideals (n) of the integers are prime. It will turn out that they are the ones in which are generated by a prime number $n = p$, and this amounts to the fundamental property of prime numbers: if a prime p divides xy, then p divides x or p divides y (Euclid's Lemma).

The following two observations are good to keep in mind; their proofs just involve matching up various definitions, so are left as exercises.

PROPOSITION A.14. *A ring is an integral domain iff the zero ideal is prime.*

PROPOSITION A.15. *A ring is a field iff the zero ideal is maximal.*

PROPOSITION A.16. *In a principal ideal domain, every nonzero prime ideal is maximal.*

PROOF. Suppose (a) is a prime ideal which is not maximal: then we have a proper containment of ideals $(a) \subsetneq (b)$, with (b) a proper ideal. By Proposition A.12, this means that $a = bc$ for some $c \in R$. Since (a) is prime, we get that either $b \in (a)$ or $c \in (a)$. The former implies that $(a) = (b)$, contradicting the strictness of the containment. So $c \in (a)$; say $c = da$. Then $a = b(d)a = bda$. Since $a \neq 0$, we can cancel, getting $bd = 1$. Thus b is a unit, so (b) is not a proper ideal, a contradiction. □

Example: In $\mathbb{Z}[T]$, the ideal (T) is prime: a polynomial is divisible by T iff its constant term is zero. And if $P_1(T)$ has constant term c_1 and $P_2(T)$ has constant term c_2, then $P_2(T)$ has constant term c_1c_2, so if $P_1(T)P_2(T)$ has constant term zero, so does at least one of P_1 and P_2. On the other hand it is not maximal, since it is strictly contained in the proper ideal $(2, T)$.

6.2. Quotient rings.
The most important use of ideals is the quotient construction: if I is an ideal in a (still assumed commutative) ring R, then we can form a ring R/I endowed with a canonical homomorphism $R \to R/I$, as follows:

The elements of R/I are the cosets $r + I$ of the subgroup I of R. The addition and multiplication laws are derived from those on R:

$$(r_1 + I) + (r_2 + I) = (r_1 + r_2 + I).$$
$$(r_1 + I) \cdot (r_2 + I) = (r_1 r_2 + I).$$

One must check that these definitions actually make sense ("are well-defined"): namely, that the sum and product of cosets does not depend upon the choice of

representative we chose. After all, $r_1 + I$ is the same coset as $r_1 + i_1 + I$, for any $i_1 \in I$. Now we just check that the properties of I are exactly such as to ensure that the final answer tolerates such ambiguity: suppose we chose $r_1 + i_1$ and $r_2 + i_2$ instead. Then we would have defined the sum to be

$$r_1 + i_1 + r_2 + i_2 + I = r_1 + r_2 + (i_1 + i_2 + I).$$

But since $i_1,\ i_2 \in I$, so is $i_1 + i_2$, which means that $i_1 + i_2 + I = I$, so it's okay: we get the same coset no matter what i_1 and i_2 we pick. And similarly for multiplication:

$$(r_1 + i_1 + I)(r_2 + i_2 + I) = (r_1 + i_1)(r_2 + i_2) + I = r_1 r_2 + r_1 i_2 + r_2 i_1 + i_1 i_2.$$

But again by the absorption property (IR2) of ideals, $r_1 i_2$, $r_2 i_1$, and $i_1 i_2$ are all elements of I, and hence so is their sum. (This shows why a mere subring wouldn't do!) Thus R/I is indeed a ring. Moreover, the map $R \to R/I$ is just $r \mapsto r + I$. It is essentially tautological that it is a homomorphism of rings.

When two elements r_1 and r_2 determine the same coset $r_1 + I = r_2 + I$, their images in R/I are equal (and conversely). In this situation, it is useful to say that r_1 and r_2 are equal *modulo* I.

Basic example: Consider the ideal (n) in $\mathbb{Z}$, where n is some positive integer. Then a choice of representative for each coset $\mathbb{Z} + (n)$ is obtained by taking $0, 1, \ldots, n-1$. In other words, for any two distinct integers $0 \leq i,\ j < n$, $i - j$ is not a multiple of n, so $i + (n)$ and $j + (n)$ are distinct cosets. Moreover, for any larger integer k, the coset $k + (n)$ will be equal to a unique coset $i + (n)$, where i is the remainder upon dividing k by n.

Note that the ring $\mathbb{Z}/(n)$ is nothing else than the finite ring we denoted Z_n, and the way in which the fact that taking usual addition and multiplication and then taking the remainder upon division by n endow Z_n with the structure of a ring is made rigorous by the quotient construction: it is a systematization of the process of "throwing away multiples of n." I highly recommend thinking about the quotient construction in this case, since all the abstract ideas are there and it is relatively easy to understand what it means for two integers a and b to determine the same cosets.

Many properties of ideals I are equivalent to certain properties of the quotient ring R/I. Here are two very important examples:

PROPOSITION A.17. *Let I be an ideal in a ring R.*
a) I is prime iff R/I is an integral domain.
b) I is maximal iff R/I is a field.

PROOF. To say that I is prime is to say that when $xy \in I$, either $x \in I$ or $y \in I$. Now an element x lies in I iff its image in R/I is zero, so an ideal I is prime if whenever a product of two elements $x + I$ and $y + I$ is zero in R/I, at least one of $x + I$ and $y + I$ is zero. This is exactly the definition of an integral domain!

If R/I is a field, then whenever x is not an element of I, there exists an element $y \in R$ such that $(x + I)(y + I) = (xy + I) = (1 + I)$, i.e., $xy - 1 = i \in I$. If I is not maximal, there exists some element $x \in R \setminus I$ and a proper ideal J containing I and x. But $1 = xy - i$, so any ideal which contains both I and x also contains xy

and $-i$, hence contains 1, so is not proper. Similarly, if I is maximal and x is any element of $R \setminus I$, then the set $I_x = \{i + rx \mid i \in I,\ r \in R\}$ is an ideal containing I and x, hence I_x strictly contains I so must contain 1. That is, $1 = i + yx$ for some $i \in I,\ y \in R$, and this means that $(x + I)(y + I) = 1 + I$, so that $x + I$ is invertible in R/I. □

EXAMPLE A.18. *Let $R = F[X, Y]$ for any field F. It is easy to check that $R/(X) \cong F[Y]$: we are considering polynomials $P(X, Y)$ modulo multiples of X, and this is amounts to evaluating $X = 0$ and considering the corresponding polynomials $P(0, Y)$, which form the ring $F[Y]$. Since the quotient ring $F[Y]$ is an integral domain, (X) is a prime ideal. Since it is not a field, (X) is not maximal. Therefore R is not a PID. Note that we showed this without exhibiting any particular nonprincipal ideal. Tracking through the preceding proofs, we see that there must be a nonprincipal ideal which contains (X); can you find one?*

APPENDIX B

More on Commutative Groups

1. Reminder on Quotient Groups

Let G be a group and H a subgroup of G. We have seen that the left cosets xH of H in G give a partition of G. Motivated by the case of quotients of rings by ideals, it is natural to consider the product operation on cosets. Recall that for any subsets S, T of G, by ST we mean $\{st \mid s \in S, t \in T\}$.

If G is commutative, the product of two left cosets is another left coset:

$$(xH)(yH) = xyHH = xyH.$$

In fact, what we really used was that for all $y \in G$, $yH = Hy$. For an arbitrary group G, this is a property of the subgroup H, called **normality**. But it is clear – and will be good enough for us – that if G is commutative, all subgroups are normal.

If G is a group and H is a normal subgroup, then the set of left cosets, denoted G/H, itself forms a group under the above product operation, called the **quotient group** of G by H. The map which assigns $x \in G$ to its coset $xH \in G/H$ is in fact a surjective group homomorphism $q : G \to G/H$, called the **quotient map** (or in common jargon, the "natural map"), and its kernel is precisely the subgroup H.

THEOREM B.1. *(Isomorphism theorem) Let $f : G \to G'$ be a surjective homomorphism of groups, with kernel K. Then G/K is isomorphic to G'.*

PROOF. We define the isomorphism $q(f) : G/K \to G'$ in terms of f: map the coset xK to $f(x) \in G'$. This is well-defined, because if $xK = x'K$, then $x' = xk$ for some $k \in K$, and then

$$f(x') = f(x)f(k) = f(x) \cdot e = f(x),$$

since k is in the kernel of f. It is immediate to check that $q(f)$ is a homomorphism of groups. Because f is surjective, for $y \in G'$ there exists $x \in G$ such that $f(x) = y$ and then $q(f)(xK) = y$, so $q(f)$ is surjective. Finally, if $q(f)(xK) = e$, then $f(x) = e$ and $x \in K$, so $xK = K$ is the identity element of G/K. □

In other words, a group G' is (isomorphic to) a quotient of a group G iff there exists a surjective group homomorphism from G to G'.

COROLLARY B.2. *If G and G' are finite groups such that there exists a surjective group homomorphism $f : G \to G'$, then $\#G' \mid \#G$.*

PROOF. $G' \cong G/\ker f$, so $\#G' \cdot \#(\ker f) = \#G$. □

Remark: Suitably interepreted, this remains true for infinite groups.

COROLLARY B.3. *If G' is isomorphic to a quotient group of G and G'' is isomorphic to a quotient group of G', then G'' is isomorphic to a quotient group of G.*

PROOF. We have surjective homomorphisms $q_1 : G \to G'$ and $q_2 : G' \to G''$, so the composition $q_2 \circ q_1$ is a surjective homomorphism from G to G''. □

2. Cyclic Groups

Recall that a group G is cyclic if there exists some element g in G such that every x in g is of the form g^n for some integer n. (Here we are using the conventions that $g^0 = e$ is the identity element of G and that $g^{-n} = (g^{-1})^n$.) Such an element g is called a generator. In general, a cyclic group will have more than one generator, and it is a number-theoretic problem to determine how many generators there are.

Example 1: The integers $\mathbb{Z}$ under addition are a cyclic group, because 1 is a generator. The only other generator is -1.

Example 2: We denote by Z_n the additive group of the ring $(\mathbb{Z}/n\mathbb{Z})$. It is also a cyclic group, because it is generated by the class of 1 (mod n).

We claim that these are the only cyclic groups, up to isomorphism. One (comparatively sophisticated) way to see this is as follows: let G be a cyclic group, with generator g. Then there is a unique homomorphism f from the additive group of the integers to G which maps 1 to g. The map f is surjective because, by assumption, every y in G is of the form g^n for some $n \in \mathbb{Z}$, i.e., $y = g^n = f(n)$. Let K be the kernel of this homomorphism. Then it is a subgroup of $(\mathbb{Z}, +)$, and since every additive subgroup of $(\mathbb{Z}, +)$ is an ideal, we have $K = n\mathbb{Z}$ for some $n \in \mathbb{N}$. Therefore by the isomorphism theorem, we have that G is isomorphic to the additive group of the quotient ring $\mathbb{Z}/n\mathbb{Z}$, i.e., to Z_n.

COROLLARY B.4. *Every quotient group of a cyclic group is cyclic.*

PROOF. We saw that a group is cyclic iff it is isomorphic to a quotient of $(\mathbb{Z}, +)$. Therefore a quotient G' of a cyclic group is a group that is isomorphic to a quotient of a quotient of $(\mathbb{Z}, +)$, and by Corollary B.3 this simply means that G' is isomorphic to a quotient of $(\mathbb{Z}, +)$ and hence is itself cyclic. □

PROPOSITION B.5. *Let $n \in \mathbb{Z}^+$. For every positive divisor k of n, there is a unique subgroup of Z_n of order k, and these are the only subgroups of Z_n.*

PROOF. For any divisor k of n, the subgroup generated by k (mod n) of $(\mathbb{Z}/n\mathbb{Z}, +)$ has order $\frac{n}{k}$, and as k runs through the positive divisors of n so does $\frac{n}{k}$. So there is at least one cyclic subgroup of Z_n of order any divisor of n. Conversely, let H be a subgroup of $(\mathbb{Z}/n\mathbb{Z}, +)$ and let k be the least positive integer such that the class of k mod n lies in H. I leave it to you to show that H is the subgroup generated by k (mod n). □

Remark: Here is a slicker proof: the subgroups of Z_n correspond to the ideals in $\mathbb{Z}/n\mathbb{Z}$ which – by a general principle on ideals in quotient rings – correspond to the ideals of $\mathbb{Z}$ containing $(n\mathbb{Z})$, which correspond to the positive divisors of n.

COROLLARY B.6. *Subgroups of cyclic groups are cyclic.*

PROPOSITION B.7. *For $a \in \mathbb{Z}^+$, the order of $a \in (\mathbb{Z}/n\mathbb{Z}, +)$ is $\frac{n}{\gcd(a,n)}$.*

PROOF. Let $d = \gcd(a, n)$ and write $a = da'$. The (additive) order of $a \pmod n$ is the least positive integer k such that $n \mid ka$. We have $n \mid ka = kda' \iff \frac{n}{d} \mid ka'$, and since $\gcd(\frac{n}{d}, a) = 1$, the least such k is $\frac{n}{d}$. □

COROLLARY B.8. *Let $a \in \mathbb{Z}$, $n \in \mathbb{Z}^+$.*
a) The class of $a \in Z_n$ is a generator if and only if $\gcd(a, n) = 1$. In particular there are $\varphi(n)$ generators.
b) For any $d \mid n$, there are precisely $\varphi(d)$ elements of Z_n of order d.
c) It follows that $\sum_{d \mid n} \varphi(d) = n$.

PROOF. Part a) is immediate from Proposition B.7. For any $d \mid n$, each element of order d generates a cyclic subgroup of order d, and we know that there is exactly one such subgroup of Z_n, so the elements of order d are precisely the $\varphi(d)$ generators of this cyclic group. Part c) follows: the left hand side gives the number of elements of order d for each $d \mid n$ and the right hand side is $\#Z_n$. □

This leads to a very useful result:

THEOREM B.9. *(Cyclicity criterion) Let G be a finite group, not assumed to be commutative. Suppose that for each $n \in \mathbb{Z}^+$, there are at most n elements x in G with $x^n = e$. Then G is cyclic.*

PROOF. Suppose G has order N, and for all $1 \leq d \leq N$, let $f(d)$ be the number of elements of G of order d. By Lagrange's Theorem, $f(d) = 0$ unless $d \mid N$, so $N = \#G = \sum_{d \mid N} f(d)$. Now, if $f(d) \neq 0$ then there exists at least one element of order d, which therefore generates a cyclic group of order d, whose elements give d solutions to the equation $x^d = e$. By our assumption there cannot be any more solutions to this equation, hence all the elements of order d are precisely the $\varphi(d)$ generators of this cyclic group. In other words, for all $d \mid n$ we have either $f(d) = 0$ or $f(d) = \varphi(d)$, so in any case we have

$$N = \sum_{d \mid N} f(d) \leq \sum_{d \mid N} \varphi(d) = N.$$

Therefore we must have $f(d) = \varphi(d)$ for all $d \mid N$, including $d = N$, i.e., there exists an element of G whose order is the order of G: G is cyclic. □

COROLLARY B.10. *Let F be a field, and let $G \subset F^\times$ be a finite subgroup of the group of nonzero elements of F under multiplication. Then G is cyclic.*

PROOF. By basic field theory, for any $d \in \mathbb{Z}^+$ the degree d polynomial $t^d - 1$ can have at most d solutions, so the hypotheses of Theorem B.9 apply to G. □

3. Products of Elements of Finite Order in a Commutative Group

Let G be a commutative group, and let $x, y \in G$ be two elements of finite order, say of orders m and n respectively. There is a unique smallest subgroup $H = H(x, y)$ of G containing both x and y, called the **subgroup generated by** x and y. $H(x, y)$ is the set of all elements of the form $x^a y^b$ for $a, b \in \mathbb{Z}$. Moreover, since x has order m and y has order n, we may write every element of H as $x^a y^b$ with $0 \leq a < m$, $0 \leq b < n$, so that $\#H \leq mn$. In particular the subgroup of an abelian group

generated by two elements of finite order is itself finite.

It is very useful to have some information about both the size of $H(x,y)$ and the order of the element xy in terms of m and n alone. However we cannot expect a complete answer:

Example 3: Suppose that $m = n = N$. We could take G to be the additive group of $\mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$, $x = (1,0)$, $y = (0,1)$. Then the subgroup generated by x and y is all of G, so has order N^2, and the order of $x+y$ is N. On the other hand, we could take $G = Z_N$ and $x = y = g$ some generator. Then $H(x,y) = G$ has order N and $\#xy$ is N if N is odd and $\frac{N}{2}$ is N is even. Or we could have taken $y = x^{-1}$ so that the $xy = e$ and has order 1. And there are yet other possibilities.

Example 4: Suppose that $\gcd(m,n) = 1$. We can show that xy has order mn, and hence is a generator for $H(x,y)$. Indeed, let $a \in \mathbb{Z}^+$ be such that $(xy)^a = e$, i.e., $x^a = y^{-a}$. But the order of x^a divides m and the order of y^{-a} divides n; since $\gcd(m,n) = 1$, $x^a = y^{-a} = 1$, so that $a \mid m,\ a \mid n$. Since, again, $\gcd(m,n) = 1$, this implies $a \mid mn$.

The general case is as follows:

THEOREM B.11. *Let x and y be elements of finite order m and n in a commutative group G. Denote by $H(x,y)$ the subgroup generated by x and y.*
a) $\operatorname{lcm}(m,n) \mid \#H(x,y) \mid mn$.
b) $\frac{\operatorname{lcm}(m,n)}{\gcd(m,n)} \mid \#(xy) \mid \operatorname{lcm}(m,n)$.

PROOF. Step 1: Define a surjective homomorphism of groups $\Psi : Z_m \times Z_n \to H(x,y)$ by $(c,d) \mapsto x^c y^{-d}$, so by $\#H(x,y) \mid \#(Z_m \times Z_n)$ by Corollary B.2.

Step 2: Let K be the kernel of Ψ. By the Isomorphism theorem, $\#H(x,y) = \#(Z_m \times Z_n)/\#K = \frac{mn}{\#K}$, so $\#H(x,y) \mid mn$. Moreover, the kernel K consists of pairs (c,d) such that $x^c = y^d$. Let $f = \gcd(m,n)$. Let o be the order of $x^c = y^d$. Since the order of x^c divides m and the order of y^d divides n, $o \mid \gcd(m,n) = f$. There are f values of $c \pmod m$ for which x^c has order dividing f, and for each of these values, there is at most one value of $d \pmod n$ such that $x^c = y^d$ (because the elements y^i for $0 \le i < n$ are distinct elements of G). This shows that the kernel can be viewed as a subset of Z_f, and it is easily checked to be a subgroup. So $\#K \mid f$ and hence

$$\operatorname{lcm}(m,n) = \frac{mn}{f} \mid \frac{mn}{\#K} = \#H(x,y).$$

Step 3: $(xy)^{\operatorname{lcm}(m,n)} = x^{\operatorname{lcm}(m,n)} y^{\operatorname{lcm}(m,n)} = 1$, so the order of xy divides $\operatorname{lcm}(m,n)$.

Step 4: Finally, suppose that $a \in \mathbb{Z}^+$ is such that $(xy)^a = x^a y^a = 1$, so $x^a = y^{-a}$. So the order of x^a, which is $\frac{m}{\gcd(a,m)}$ is equal to the order of y^{-a}, which is $\frac{n}{\gcd(a,n)}$. In other words, we have

$$m \gcd(a,n) = n \gcd(a,m).$$

Since $\gcd(\frac{m}{f}, n) = 1$, $\frac{m}{f} \mid \gcd(a,m)$, or

$$m \mid f \gcd(a,m) \mid fa.$$

Similarly

$$n \mid f \gcd(a, n) \mid fa.$$

Therefore $\operatorname{lcm}(m, n) \mid fa$, or $\frac{\operatorname{lcm}(m,n)}{\gcd(m,n)} \mid a$, completing the proof of the theorem. □

Remark: The divisibilities in Theorem B.11 are best possible: if h and o are positive integers such that $\operatorname{lcm}(m, n) \mid h \mid mn$ and $\frac{\operatorname{lcm}(m,n)}{\gcd(m,n)} \mid o \mid \operatorname{lcm}(m, n)$, then there exist elements $x, y \in Z_m \times Z_n$ such that $\#H(x, y) = h$, $\#xy = o$.

Remark: The situation is profoundly different for noncommutative groups: for every $m, n \geq 2$ and $2 \leq r \leq \infty$ there exists a group G containing elements x of order m, y of order n whose product xy has order r. Moreover, if $r < \infty$ then one can find a finite group G with these properties, whereas one can find an infinite group with these properties iff $\frac{1}{m} + \frac{1}{n} + \frac{1}{r} \leq 1$.

The following is a consequence of Theorem B.11 (but is much simpler to prove):

COROLLARY B.12. *Let $m, n \in \mathbb{Z}^+$. Then $Z_m \times Z_n$ is cyclic iff $\gcd(m, n) = 1$.*

PROOF. The order of any element (c, d) divides $\operatorname{lcm}(m, n)$, and the order of $(1, 1)$ is $\operatorname{lcm}(m, n)$. So the group is cyclic iff $mn = \operatorname{lcm}(m, n)$ iff $\gcd(m, n) = 1$. □

4. Character Theory of Finite Abelian Groups

4.1. Introduction.

In this section our goal is to present the theory of characters of finite abelian groups. Although this is an "easy" theory in that we can present it in its entirety here, it nevertheless of the highest impotance, being the jumping off point for at least two entire disciplines of mathematics: the general theory of linear representations of groups, and Fourier analysis. The special case of characters of the unit groups $U(N) = (\mathbb{Z}/N\mathbb{Z})^\times$ will be used as one of the essential ingredients in the proof of Dirichlet's theorem on primes in arithmetic progessions.

Let G be a finite commutative group. A **character** $\chi : G \to \mathbb{C}^\times$ of G is a homomorphism from G to the group $\mathbb{C}^\times$ of nonzero complex numbers under multiplication.

Suppose $N = \#G$. By Lagrange's theorem we have, for any $g \in G$, that $g^N = e$ (the identity element), and thus for any character χ on G we have

$$\chi(g)^N = \chi(g^N) = \chi(e) = 1.$$

Thus $\chi(g)$ is itself a complex Nth root of unity. Recall that the set of all complex Nth roots of unity forms a cyclic group of order N, say μ_N. In other words, every character on a group G of order N is really just a homomorphism from G to μ_N, or equally well, from G into any fixed order N cyclic group.

We write $X(G)$ for the set of all characters of G. We can endow $X(G)$ with the structure of a group: given $\chi_1,\ \chi_2 \in X(G)$, we define their product "pointwise":

$$\forall g \in G,\ (\chi_1\chi_2)(g) := \chi_1(g)\chi_2(g).$$

The identity element is the **trivial character** $g \mapsto 1$ for all g, and the inverse of χ is the function $\chi^{-1} : g \mapsto \frac{1}{\chi(g)}$. Because for any $z \in \mathbb{C}$ we have $z\overline{z} = |z|^2$, if if z is a root of unity, then the inverse of z is given by its complex conjugate $\overline{z}$. It follows that the inverse of a character χ is also given by taking complex conjugates:

$$\overline{\chi}(g) = \overline{\chi(g)} = \frac{1}{\chi(g)} = \chi^{-1}(g).$$

4.2. The Character Extension Lemma.

Most of the content of the entire theory resides in the following result.

LEMMA B.13. *(Character Extension Lemma) Let H be a subgroup of a finite commutative group G. For any character $\psi : H \to \mathbb{C}^\times$, there are precisely $[G : H]$ characters $\Psi : G \to \mathbb{C}^\times$ such that $\Psi|_H = \psi$.*

PROOF. The result is clear if $H = G$, so we may assume there is $g \in G \setminus H$. Let $H_g = \langle g, H\rangle$ be the subgroup generated by H and g. Now we may or may not have $H_g = G$, but suppose that we can establish the result for the group H_g and its subgroup H. Then the general case follows by induction, since for any $H \subset G$ choose $g_1, \ldots, g_n$ such that $G = \langle H, g_1, \ldots, g_n\rangle$. Then we can define $G_0 = H$ and for $1 \leq i \leq n$, $G_i = \langle G_{i-1}, g_i\rangle$. Applying the Lemma in turn to G_{i-1} as a subgroup of G_i gives that in all the number of ways to extend the character ψ of $H = G_0$ is

$$[G_1 : G_0][G_2 : G_1] \cdots [G_n : G_{n-1}] = [G : G_0] = [G : H].$$

So let us now prove that the number of ways to extend ψ from H to $H_g = \langle H, g\rangle$ is $[H_g : H]$. For this, let d be the order of g in G, and consider $\tilde{G} := H \times \langle g\rangle$. The number of ways to extend a character ψ of H to a character of $\tilde{G}$ is equal to $\#\langle g\rangle = d$: such a homomorphism is uniquely specified by the image of $(1, g)$ in $\mu_d \subset \mathbb{C}^\times$, and all d such choices give rise to homomorphisms.

Moreover, there is a surjective homomorphism $\varphi : H \times \langle g\rangle$ to H_g: we just take $(h, g^i) \mapsto hg^{-i}$. The kernel of φ is the set of all pairs (h, g^i) such that $g^i = h$. In other words it is precisely the intersection $H \cap \langle g\rangle$, which has cardinality a divisor of d, say e. It follows that

$$\#H_g = \frac{\#H \times \langle g\rangle}{\#H \cap \langle g\rangle} = \frac{d}{e} \cdot \#H,$$

so

$$[H_g : H] = \frac{d}{e}.$$

But a homomorphism $f : H \times \langle g\rangle \to \mathbb{C}^\times$ descends to a homomorhpism on the quotient H_g iff it is trivial on the kernel of the quotient map, i.e., is trivial on $H \cap \langle g\rangle$. In other words, the extensions of ψ to a character of H_g correspond precisely to the number of ways to map the order d element g into $\mathbb{C}^\times$ such that $g^{\frac{d}{e}}$ gets mapped to 1. Thus we must map g to a $(\frac{d}{e})$th root of unity, and conversely all such mappings induce extensions of ψ. Thus the number of extensions is $\frac{d}{e} = [H_g : H]$. □

COROLLARY B.14. *For any finite commutative group G, $X(G)$ is finite and*

$$\#X(G) = \#G.$$

PROOF. Apply Lemma B.13 with $H = 1$. □

COROLLARY B.15. *For G a finite commmutative group and $g \in G$, TFAE:*
(i) For every $\chi \in X(G)$, $\chi(g) = 1$.
(ii) g is the identity element e of G.

PROOF. Certainly (ii) $\implies$ (i). Conversely, if $g \neq e$, then $H := \langle g \rangle$ is a nontrivial cyclic group. By Corollary B.14, there exists a nontrivial character ψ of H. Since g generates H, this implies $\psi(g) \neq 1$. Now apply Lemma B.13 to extend ψ to a character of G. □

From these results one can deduce that the character group construction behaves nicely under homomorphisms: suppose $f : G \to H$ is a homomorphism of finite commutative groups. Then we can define a map $X(f) : X(H) \to X(G)$ – note well: in the opposite direction! – just by taking a character $\chi : H \to \mathbb{C}^\times$ and precomposing it with f to get a character $\chi \circ f : G \to \mathbb{C}^\times$.

PROPOSITION B.16. *Let $f : G \to H$ be a homomorphism of finite commutative groups.*
a) The induced map $X(f) : X(H) \to X(G)$ is a group homomorphism.
b) The homomorphism f is injective $\iff$ the homomorphism $X(f)$ is surjective.
c) The homomorphism f is surjective $\iff$ the homomorphism $X(f)$ is injective.

PROOF. Part a) is a straightforward verification which we leave to the reader.

b) Assume first that f is injective. We may as well assume then that G is a subgroup of H and $f = \iota$ is the inclusion map. Then the induced homomorphism $X(\iota) : X(H) \to X(G)$ is nothing else than the map which restricts a character of H to a character of the subgroup G; that this restriction map is surjective is an immediate consequence of Lemma B.13. Inversely, assume that f is not injective, so that there exists $e \neq g \in G$ such that $f(g) = e \in H$. By Corollary B.15, there exists a character $\chi : G \to \mathbb{C}^\times$ such that $\chi(g) \neq 1$. But then for any character $\psi : H \to \mathbb{C}^\times$, we have

$$(\psi \circ f)(g) = \psi(e) = 1,$$

which shows that $\psi \circ f \neq \chi$, i.e., χ is not in the image of $X(f)$.

c) By the Character Extension Lemma, there are precisely $[H : f(G)]$ characters on H which are trivial on $f(G)$. Therefore f is surjective iff a character ψ on H for which $\psi \circ f$ is trivial is necessarily itself trivial. □

4.3. Orthogonality relations.

THEOREM B.17. *Let G be a finite abelian group, with character group G.*
a) For any nontrivial character $\chi \in X(G)$, we have $\sum_{g \in G} \chi(g) = 0$.
b) For any nontrivial element g of G, we have $\sum_{\chi \in X(G)} \chi(g) = 0$.

PROOF. a) Put

$$S = \sum_{g \in G} \chi(g). \tag{66}$$

Since χ is nontrivial, there exists $g_0 \in G$ such that $\chi(g_0) \neq 1$. Multiplying both sides of (66) by $\chi(g_0)$, we get

$$\chi(g_0)S = \sum_{g \in G} \chi(g)\chi(g_0) = \sum_{g \in G} \chi(gg_0) = \sum_{g \in G} \chi(g) = S;$$

the penultimate equality holds because, as g runs through all elements of G, so does g_0. Therefore we have

$$(\chi(g_0) - 1)\, S = 0.$$

Since $\chi(g_0) \neq 1$, we must have $S = 0$.
b) If $g \neq e$, then by Corollary B.15 there is a character χ such that $\chi(g) \neq 1$, and then the argument is identical to part a).[1] □

Let us explain why these are called orthogonality relations. Consider the set $\mathbb{C}^G$ of all functions $f : G \to \mathbb{C}$. Under pointwise addition and scalar multiplication, $\mathbb{C}^G$ is a $\mathbb{C}$-vector space of dimension $\#G$. We define a Hermitian inner product on $\mathbb{C}^G$ by

$$\langle f, g \rangle := \frac{1}{\#G} \sum_{x \in G} f(x)\overline{g(x)}.$$

Now let χ_1 and χ_2 be characters of G. If $\chi_1 = \chi_2$, then we have

$$\langle \chi_1, \chi_1 \rangle = \frac{1}{\#G} \sum_{x \in G} |\chi_1(x)|^2 = 1,$$

whereas if $\chi_1 \neq \chi_2$, then $\chi_1 \chi_2^{-1}$ is nontrivial, and then Theorem B.17 gives

$$\langle \chi_1, \chi_2 \rangle = \frac{1}{\#G} \sum_{x \in G} (\chi_1 \chi_2^{-1})(x) = 0.$$

In other words, the set $X(G)$ of characters of G is orthonormal with respect to the given inner product. In particular, the subset $X(G)$ of $\mathbb{C}^G$ is linearly independent. Since its cardinality, $\#G$, is equal to the dimension of $\mathbb{C}^G$, we conclude:

COROLLARY B.18. *Let G be a finite commutative group, and let $\mathbb{C}^G$ be the $\mathbb{C}$-vector space of all functions from G to $\mathbb{C}$, endowed with the inner product*

$$\langle f, g \rangle = \frac{1}{\#G} \sum_{x \in G} f(x)\overline{g(x)}.$$

Then the set of characters of G forms an orthonormal basis with respect to $\langle\ ,\ \rangle$. Therefore, any function $f : G \to \mathbb{C}$ can be expressed as a unique linear combination of characters. Explicitly:

$$f = \sum_{\chi \in X(G)} \langle f, \chi \rangle \chi.$$

This can be viewed as the simplest possible case of a **Fourier inversion formula**.

4.4. The canonical and illicit isomorphism theorems; Pontrjagin duality.

In the course of study of finite commutative groups, one sees that subgroups and quotient groups have many similar properties. For instance, subgroups of cyclic groups are cyclic, and also quotients of cyclic groups are cyclic. Moreover, a cyclic group of order n has a unique subgroup of every order dividing n and no other subgroups, and the same is true for its quotients. If one plays around for a bit with finite commutative groups, one eventually suspects the following result:

[1] Alternately, using the canonical isomorphism $G \cong X(X(G))$ described in the next section, one can literally deduce part b) from part a).

THEOREM B.19. *Let G and H be finite commutative groups. Then TFAE:*
(i) H can be realized as a subgroup of G: $\exists$ an injective homomorphism $H \to G$.
(ii) H can be realized as a quotient of G: $\exists$ a surjective homomorphism $G \to H$.

There is a certain resemblance between Theorem B.19 and Proposition B.16, but they are not the same. Proposition B.16 asserts that if there is an injection $H \to G$, there is a surjection $X(G) \to X(H)$ (and similarly with "injection" and "surjection" interchanged). To deduce Theorem B.19 from Proposition B.16, one needs the following:

THEOREM B.20. *(Illicit Isomorphism Theorem) Any finite commutative group G is isomorphic to its chracter group $X(G)$.*

Some cases of Theorem B.20 are easy to establish. For instance, since G and $X(G)$ have the same order, they must be isomorphic whenever $\#G$ is prime. Further, to give a character on a cyclic group of order N it suffices to send a fixed generator to any Nth root of unity in $\mathbb{C}$. More precisely, choosing a generator of an abstract cyclic group G order N amounts to choosing an isomorphism of G with $\mathbb{Z}/N\mathbb{Z}$ (we send the generator to 1 (mod N)). And the characters on $\mathbb{Z}/N\mathbb{Z}$ are all obtained by exponentiation: for any $c \in \mathbb{Z}/N\mathbb{Z}$, there is a unique character χ_a such that

$$\chi_c(1) = e^{2\pi ic/N}$$

and therefore for any $b \in \mathbb{Z}/N\mathbb{Z}$

$$\chi_c(b) = e^{2\pi icb/N}.$$

It is immediate to check that $\chi_c \cdot \chi_{c'} = \chi_{c+c'}$, where addition is taken mod N. Thus we get a canonical isomorphism $X(\mathbb{Z}/N\mathbb{Z}) \xrightarrow{\sim} \mathbb{Z}/N\mathbb{Z}$.

Moreover, if G_1 and G_2 are finite commutative groups, then in a natural way

$$X(G_1 \times G_2) = X(G_1) \times X(G_2);$$

again we leave the details to the interested reader. Of course the analogous identity for products of any finite number of groups follows by induction.

Combining these observations, it follows that $G \cong X(G)$ for any finite commutative group G of the form $Z_{n_1} \times \ldots \times Z_{n_k}$, i.e., for any direct product of cyclic groups. Is this enough to prove Theorem B.20? Indeed it is, because of the following:

THEOREM B.21. *(Fundamental theorem on finite commutative groups) Let G be a finite commutative group.*
a) There exist prime powers $p_1^{a_1}, \ldots, p_r^{a_r}$ (we allow $p_i = p_j$ for $i \neq j$) such that

$$G \cong Z_{p_1^{a_1}} \times \ldots \times Z_{p_r^{a_r}},$$

i.e., G is a direct product of finite cyclic groups of prime power order.
b) Moreover, this decomposition is essentially unique in the following (familiar) sense: if also we have

$$G \cong Z_{q_1^{b_1}} \times \ldots \times Z_{q_s^{b_s}},$$

then $r = s$ and there exists a bijection $\sigma : \{1, \ldots, r\} \to \{1 \ldots s\}$ such that for all $1 \leq i \leq r$, $q_{\sigma(i)} = p_i$ and $b_{\sigma(i)} = a_i$.

Now please bear with me while I make a few possibly confusing remarks about why I have labelled Theorem B.20 the "illicit" isomorphism theorem. In some sense it is "lucky" that $G \cong X(G)$, in that it is not part of the general meaning of "duality" that an object be isomorphic to its dual object. Rather, what one has in much more generality is a canonical injection from an object to its **double dual**. Here, this means the following: we can construct a canonical map $G \to X(X(G))$. In other words, given an element g in G, we want to define a character, say $g\bullet$, *on* the character group, i.e., a homomorphism $X(G) \to \mathbb{C}^\times$. This may sound complicated at first, but in fact there is a very easy way to do this: define $g \bullet \chi := \chi(g)$! It is no problem to check that the association $g \mapsto g\bullet$ is a homomorphism of finite abelian groups. Moreover, suppose that for any fixed $g \in G$ the map $g\bullet$ were trivial: that means that for all $\chi \in X(G)$, $\chi(g) = 1$. Applying Corollary B.15, we get that $g = 1$. Therefore this map

$$\bullet : G \to X(X(G))$$

is an injective homomorphism between finite abelian groups. Moreover,

$$\#X(X(G)) = \#X(G) = \#G,$$

so it is an injective homomorphism between finite groups of the same order, and therefore it must be an isomorphism.

In order to write down the isomorphism $\bullet$, we did not have to make any choices. There is a precise sense in which the isomorphism to the double dual is "canonical" and any isomorphism between G and $X(G)$ is "noncanonical", but explanining this involves the use of category theory so is not appropriate here. More interesting is to remark that there is a vastly more general class of commutative groups G for which $X(G)$ is defined in such a way as to render true all of the results we have proved here *except* the illicit isomorphism theorem: we need not have $G \cong X(G)$. For this we take G to be a commutative group endowed with a topology which makes it locally compact Hausdorff. Any commutative group G can be endowed with the discrete topology, which gives many examples. For a finite group the discrete topology is the only Hausdorff topology, so this is certainly the right choice, but an infinite group may or may not carry other interesting locally compact topologies. Some examples:

Example 1: The integers $\mathbb{Z}$: here we do want the discrete topology.

Example 2: The additive group $(\mathbb{R}, +)$ with its usual Euclidean topology: this is a locally compact group which is neither discrete nor compact. More generally, one can take $(\mathbb{R}^n, +)$ (and in fact, if G_1 and G_2 are any two locally compact commutative groups, then so is $G_1 \times G_2$ when endowed with the product topology).

Example 3: The multiplicative group $\mathbb{C}^\times$ of the complex numbers is again locally compact but neither discrete nor compact, but it is "closer to being compact" then the additive group $\mathbb{C} \cong \mathbb{R}^2$. In fact, considering polar coordinates gives an isomorphism of topological groups $\mathbb{C}^\times \cong \mathbb{R}^{>0} \times S^1$, where S^1 is the unit circle. Moreover, the logarithm function shows that $\mathbb{R}^{>0}$ is isomorphic as a topological group to $(\mathbb{R}, +)$, so all in all $\mathbb{C}^\times \cong (\mathbb{R}, +) \times S^1$. Note that S^1, the circle group, is itself a very interesting example.

Now, given any locally compact commutative group G, one defines the **Pontrjagin dual group** $X(G)$, which is the group of all continuous group homomorphisms from G to the circle group S^1. Moreover, $X(G)$ can be endowed with a natural topology.[2] Again, one has a natural map $G \to X(X(G))$ which turns out to be an isomorphism in all cases.

If G is a finite, discrete commutative group, then as we saw, any homomorphism to $\mathbb{C}^\times$ lands in S^1 (and indeed, the countable subgroup of S^1 consisting of all roots of unity) anyway; moreover, by discreteness every homomorphism is continuous. Thus $X(G)$ in this new sense agrees with the character group we have defined. But for infinite groups Pontrjagin duality is much more interesting: it turns out that G is compact iff $X(G)$ is discrete.[3] Since a topological space is both compact and discrete iff it is finite, we conclude that a topological group G which is infinite and either discrete or compact cannot be isomorphic to its Pontrjagin dual.

It is easy to see that $\mathrm{Hom}(\mathbb{Z}, S^1) = S^1$, which according to the general theory implies $\mathrm{Hom}(S^1, S^1) = \mathbb{Z}$: the discrete group $\mathbb{Z}$ and the compact circle group S^1 are mutually dual. This is the theoretical underpinning of Fourier series.

However, if G is neither discrete nor compact, then the same holds for $X(G)$, so there is at least a fighting chance for G to be isomorphic to $X(G)$. Indeed this happens for $\mathbb{R}$: $\mathrm{Hom}(\mathbb{R}, S^1) = \mathbb{R}$, where we send $x \in \mathbb{R}$ to the character $t \mapsto e^{2\pi i t x}$. This is the theoretical underpinning of the Fourier transform.

Another sense in which the isomorphism between G and $X(G)$ for a finite commutative group G is "illicit" is that turns out not to be necessary in the standard number-theoretic applications. A perusal of elementary number theory texts reveals that careful authors take it as a sort of badge of honor to avoid using the illicit isomorphism, even if it makes the proofs a bit longer. For example, the most natural analysis of the group structure of $(\mathbb{Z}/2^a\mathbb{Z})^\times$ for $a \geq 3$ would consist in showing: (i) the group has order 2^{a-1}; (ii) it has a cyclic subgroup of order 2^{a-2}; (iii) it has a noncyclic quotient so is itself not cyclic. Applying Theorem B.21 one can immediately conclude that it must be isomorphic to $Z_{2^{a-2}} \times Z_2$. In our work in Handout 9.5, however, we show the isomorphism by direct means.

This was first drawn to my attention by a close reading of J.-P. Serre's text [**Se73**] in which the illicit isomorphism is never used. Following Serre, our main application of character groups – namely the proof of Dirichlet's theorem on primes in arihtmetic progressions – uses only $\#X(G) = \#G$, but not $X(G) \cong G$.

However, to my mind, avoiding the proof of Theorem B.21 gives a misleading impression of the difficulty of the result.[4] On the other hand, Theorem B.21 evidently has some commonalities with the fundamental theorem of arithmetic, which makes

[2]If you happen to know something about topologies on spaces of functions, then you know that there is one particular topology that always has nice properties, namely the **compact-open** topology. That is indeed the correct topology here.

[3]Similarly, G is discrete iff $X(G)$ is compact; this follows from the previous statement together with $G \cong X(X(G))$.

[4]The real reason it is often omitted in such treatments is that the authors know that they will be giving a more general treatment of the structure theory finitely generated modules over a principal ideal domain, of which the theory of finite commutative groups is a very special case.

it somewhat desirable to see the proof. In the next section we provide such a proof, which is not in any sense required reading.

5. Proof of the Fundamental Theorem on Finite Commutative Groups

First some terminology: Let G be a commutative group, written multiplicatively.

If $\#G = p^a$ is a prime power, we say G is a **p-group**.

For $n \in \mathbb{Z}^+$, we put $G[n] = \{x \in G \mid x^n = 1\}$. This is a subgroup of G.

We say that two $H_1,\ H_2$ subgroups of G are **complementary** if $H_1 \cap H_2 = \{1\}$, $H_1H_2 = G$. In other words, every element g of G can be uniquely expressed in the form h_1h_2, with $h_i \in H_i$. In yet *other* (equivalent) words, this means precisely that the homomorphism $H_1 \times H_2 \to G$, $(h_1, h_2) \mapsto h_1h_2$ is an isomorphism. We say that a subgroup H is a **direct factor** of G if there exists H' such that H, H' are complementary subgroups. Thus, in order to prove part a) it suffices to show that every finite commutative group has a nontrivial direct factor which is cyclic of prime power order; and in order to prove part b) it suffices (but is much harder!) to show that if $G \cong H \times H' \cong H \times H''$ then $H' \cong H''$.

More generally if we have a finite set $\{H_1, \ldots, H_r\}$ of subgroups of G such that $H_i \cap H_j = \{1\}$ for all $i \neq j$ and $G = H_1 \cdots H_r$, we say that the H_i's form a set of complementary subgroups and that each H_i is a direct factor. In such a circumstance we have $G \cong H_1 \times \ldots \times H_r$.

We now begin the proof of Theorem B.21.

Step 1 (primary decomposition): For any commutative group G, let G_p be the set of elements of G whose order is a power of p. Also let $G^{p'}$ be the set of elements of G whose order is prime to p. It follows from Theorem B.11b) that G_p and $G^{p'}$ are both subgroups of G. We claim that G_p and $G^{p'}$ are complementary subgroups. Certainly $G_p \cap G^{p'} = \{e\}$, since any element of the intersection would have both order a power of p and relatively prime to p and thus have order 1 and be the identity. On the other hand, let x be any element of G, and write its order as $p^k \cdot b$ with $\gcd(p, b) = 1$. Thus we can choose i and j such that $ip^k + jb = 1$, and then $x = x^1 = x^{ip^k+jb} = (x^{p^k})^i \cdot (x^b)^j$, and by Proposition B.7 the order of $(x^{p^k})^i$ divides b (so is prime to p) and the order of $(x^b)^j$ divides p^k. This proves the claim. Now a simple induction argument gives the following:

PROPOSITION B.22. *Let G be a finite abelian group, of order $n = p_1^{a_1} \cdots p_r^{a_r}$. Then $\{G_{p_i}\}_{i=1}^r$ forms a set of complementary subgroups, and the canonical map $H_1 \times \ldots \times H_r \to G$, $(h_1, \ldots, h_r) \mapsto h_1 \cdots h_r$ is an isomorphism.*

Thus any finite commutative group can be decomposed, in a unique way, into a direct product of finite commutative groups of prime power order. We may therefore assume that G is a commutative p-group from now on.

Step 2: We prove a refinement of Theorem B.9 for commutative p-groups.

PROPOSITION B.23. *Let p be a prime and G be a finite commutative group of order p^a for some $a \in \mathbb{Z}^+$. TFAE:*
(i) G has exactly p elements of order p.
(ii) G is cyclic.

PROOF. We already know that (ii) $\implies$ (i), of course. Assume (i); the natural strategy is to appeal to our cyclicity criterion Theorem B.9. In this case we wish to show that for any $0 < k \leq a$, there are at most p^k elements of G of order dividing p^k. We accomplish this by induction (!); the case of $k = 1$ is our hypothesis, so assume that for all $1 \leq \ell < k$ the number of elements of order dividing p^ℓ in G is at most p^ℓ and we wish to show that the number of element of order dividing p^k is at most p^k. For this, consider the endomorphism

$$\varphi : G[p^k] \to G[p^k],\ x \mapsto x^{p^{k-1}}.$$

Now the kernel of φ is precisely $G[p^{k-1}]$, which we have inductively assumed has order at most p^{k-1}. If the order of $G[p^k]$ exceeds p^k, then since

$$\varphi(G[p^k]) \cong G[p^k]/\operatorname{Ker}(\varphi),$$

we would have $\#\varphi(G[p^k]) > p$. But by Proposition B.7 the image of φ consists entirely of elements of order dividing p, contradiction. □

Step 3:

PROPOSITION B.24. *Let G be a finite commutative p-group, and let p^a be the maximum order of an element of G. Then every cyclic subgroup C of order p^a is a direct factor of G: there exists a complementary subgroup H, giving an isomorphism $G \cong C \times H$.*

PROOF. The result holds vacuously for commutative groups of order p. Assume that it holds for all commutative groups of order p^k for $k < a$, and suppose we have $G = p^a$, x an element of maximal order in G and $C = \langle x \rangle$ If the order of x is p^a, then $G = C$ is cyclic and the conclusion again holds trivially. Otherwise, by Proposition B.23, there exists an order p subgroup K of G not contained in C, so $C \cap K = \{e\}$. Then the cyclic subgroup $(C + K)/K$ has maximal order in G/K; by induction there exists a complementary subgroup H of G/K, i.e., a subgroup H containing K such that $(C + K) \cap H = K$, $(C + K) \cdot H = G$. It follows that $H \cap C \subset K \cap C = \{e\}$ and $C \cdot H = G$, so C and H are complementary subgroups. □

We may now deduce Theorem B.21a) from Proposition B.24. Indeed, given any finite p-group G we choose an element x of maximum order p^a, which generates a cyclic subgroup C of maximum order, which according to Proposition B.24 has a complementary subgroup H and thus $G \cong Z_{p^r} \cong H$. Applying the same procedure to H, eventually we will express G as a product of finite cyclic groups of p-power order.

Step 4: Finally we address the uniqueness of the decomposition of a commutative p-group into a direct product of cyclic groups.[5] Suppose we have

$$G \cong Z_{p^{a_1}} \times \ldots \times Z_{p^{a_r}} \cong Z_{p^{b_1}} \times \ldots \times Z_{p^{b_s}}.$$

[5]This part of the proof follows [**Su95**].

We may assume that $a_1 \geq \ldots \geq a_r$ and $b_1 \geq \ldots \geq b_s$, and we wish to prove that $r = s$ and $a_i = b_i$ for all i. We may also inductively assume the uniqueness statement for commutative p-groups of smaller order than G. Now let $\varphi : G \to G$ be $x \mapsto x^p$. Then we have

$$\varphi(G) \cong Z_{p^{a_1-1}} \times \ldots \times Z_{p^{a_r-1}} \cong Z_{p^{b_1-1}} \times \ldots \times Z_{p^{b_s-1}}.$$

Since $\#\varphi(G) < \#G$, by induction the two decompositions are unique, the only proviso being that if an exponent c_i is equal to 1, then $Z_{p^{c_i-1}}$ is the trivial group, which we do not allow in a direct factor decomposition. Therefore suppose that k is such that $a_i = 1$ for all $i > k$ and l is such that $b_j = 1$ for all $j > l$. Then we get $k = l$ and $a_i = b_i$ for all $1 \leq i \leq k$. But now we have

$$p^{r-k} = \frac{\#G}{p^{a_1+\ldots+a_k}} = \frac{\#G}{p^{b_1+\ldots+b_k}} = p^{s-k},$$

so we conclude $r = s$ and thus $a_i = b_i$ for $1 \leq i \leq r$.

It is interesting to ask which of the steps go through for a group which is infinite, non-commutative or both.

Step 1 fails in a non-commutative group: the elements of p-power order need not form a subgroup. For instance, the symmetric group S_n is generated by transpositions. In any commutative group one can define the subgroups G_p for primes p, and they are always pairwise disjoint. The subgroup they generate is called the **torsion subgroup** of G and often denoted $G[\text{tors}]$: it consists of all elements of finite order.

Step 2 fails for noncommutative finite p-groups: The quaternion group $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ is a noncyclic group of order $8 = 2^3 = p^3$ which has exactly $p = 2$ elements of order dividing p. It is false for all infinite abelian groups, since an infinite group can only be cyclic if its torsion subgroup is trivial.

Step 3 fails for finite noncommutative groups: again Q_8 is a counterexample.

As for Step 4, one may ask the following

QUESTION 9. *Suppose we have three groups H, G_1, G_2 such that $H \times G_1 \cong H \times G_2$. Must it then be the case that $G_1 \cong G_2$?*

Without any restrictions the answer to this question is negative. For instance, one can take $H = G_1 = (\mathbb{R}, +)$, $G_2 = 0$, and note that $\mathbb{R} \times \mathbb{R} \cong \mathbb{R}$ as $\mathbb{Q}$-vector spaces, hence as commutative groups. On the other hand:

THEOREM B.25. *(Remak-Krull-Schmidt) If H, G_1 and G_2 are all finite groups, then indeed $H \times G_1 \cong H \times G_2$ implies $G_1 \cong G_2$.*

A group G is **indecomposable** if it is not isomorphic to $H_1 \times H_2$ with H_1 and H_2 both nonzero. By Theorem B.21, a finite commutative group is indecomposable iff it is cyclic of prime power order. Any finite group can be written as a product of indecomposable groups. Using Theorem B.25 it can be shown that if

$$G \cong H_1 \times \ldots \times H_r = K_1 \times \ldots \times K_s,$$

where each H_i and K_j are indecomposable (nontrivial) groups, then $r = s$ and there exists a bijection $\sigma : \{1, \ldots, r\} \to \{1, \ldots, r\}$ such that $K_i \cong H_{\sigma(i)}$ for $1 \leq i \leq r$.

6. Wilson's Theorem in a Finite Commutative Group

Here is one of the classic theorems of elementary number theory.

THEOREM B.26. *(Wilson's Theorem) For an odd prime p, $(p-1)! \equiv -1 \pmod{p}$.*

Remark: The *converse* of Wilson's Theorem also holds: if for some integer $n > 1$ we have $(n-1)! \equiv 1 \pmod{n}$, then n is prime. In fact it can be shown that for all composite $n > 4$, $n \mid (n-1)!$ (exercise).

Most of the standard proofs involve starting with an elementary group-theoretic fact and then recasting it to avoid group-theoretic language to a greater or lesser extent. Since this handout is meant to be a "comprehensive" guide to finite commutative groups, we may as well give the argument in its proper language.

For a finite group G, let $d_2(G)$ be the number of order 2 elements in G.

THEOREM B.27. *(Wilson's Theorem in a Finite Commutative Group) Let $(G,+)$ be a finite commutative group, and let $S = \sum_{x \in G} x$. Then:*
a) If $d_2(G) \neq 1$, then $S = 0$.
b) If $d_2(G) = 1$ – so that G has a unique element, say t, of order 2 – then $d_2(G) = t$.

PROOF. We set

$$G[2] = \{x \in G \mid 2x = 0\}.$$

Every nonzero element of $G[2]$ has order 2, so by Theorem B.21, $G[2] \cong Z_2 \times \ldots \times Z_2 = Z_2^k$, a direct product of copies of the cyclic group of order 2.[6]

Consider the *involution* $\iota : G \to G$ given by $x \mapsto -x$. The *fixed points* of ι – i.e., the elements $x \in G$ such that $\iota(x) = x$ – are precisely the elements of $G[2]$. Thus the elements of $G \setminus G[2]$ occur in pairs of distinct elements $x, -x$, so $\sum_{x \in G \setminus G[2]} x = 0$. In other words, $\sum_{x \in G} x = \sum_{x \in G[2]} x$, and we are reduced to the case $G[2] \cong Z_2^k$.
Case 1: $k = 0$, i.e., $G[2] = 0$. Then

$$\sum_{x \in G[2]} x = \sum_{x \in \{0\}} x = 0.$$

Moreover, in this case $d_2(G) = 0$, in agreement with the statement of the theorem.
Case 2: $k = 1$, i.e., $G[2] = Z_2$. Then

$$\sum_{x \in G[2]} x = \sum_{x \in Z_2} x = 0 + 1 = 1,$$

where 1 is the unique element of order 2 in $Z_2 \cong G[2]$ (and thus also the unique element of order 2 in G). Again, this agrees with the statement of the theorem.
Case 3: $k \geq 2$. Then $d_2(G) \geq 3$, so we wish to show $S = \sum_{x \in Z_2^k} x = 0$. For each $1 \leq i \leq k$, half of the elements of Z_2^k have ith coordinate $0 \in Z_2$; the other half have ith coordinate $1 \in Z_2$. So the sum of the ith coordinates of the elements of Z_2^k is $2^k/2 = 2^{k-1} = 0 \in Z_2$, since $k \geq 2$: every coordinate of S equals 0, so $S = 0$. □

[6]Invocation of Theorem B.21 is overkill here: any 2-torsion commutative group admits the unique structure of a vector space over the field $\mathbb{F}_2$ with 2 elements. Being finite, $G[2]$ is certainly finite-dimensional over $\mathbb{F}_2$, so is isomorphic as a vector space – hence *a fortiori* as an additive group – to $\mathbb{F}_2^n$.

LEMMA B.28. *a) Let $G_1, \ldots, G_r$ be commutative groups. Then*

$$\left(\prod_{i=1}^r G_i\right)[2] = \prod_{i=1}^r G_i[2].$$

b) If $G_1, \ldots, G_r$ are finite, then $d_2(\prod_{i=1}^r G_i) = \prod_{i=1}^r d_2(G_i)$.

PROOF. a) More generally, consider any $x = (x_1, \ldots, x_n) \in \prod_{i=1}^r G_i$. Then the order of x is the lcm of the orders of the components x_i. Further, the lcm of a finite set of numbers divides 2 iff each number in the set divides 2.
b) This follows immediately from part a). □

We now show that Theorem B.27 implies Theorem B.26. Let $\mathbb{F}$ be a finite field. We take $G = \mathbb{F}^\times$, the multiplicative group of nonzero elements of $\mathbb{F}$.[7] Now $x \in G[2] \iff x^2 = 1$, and the polynomial $t^2 - 1$ has exactly two roots in any field of characteristic different from 2 and exactly one root in any field of characteristic 2. So $d_2(\mathbb{F}^\times)$ is equal to 1 if $\#F$ is odd and equal to 0 if $\#F$ is even. Thus:

COROLLARY B.29. *Let $\mathbb{F}$ be a finite field, put $P = \prod_{x \in \mathbb{F}^\times} x$. Then:*
a) If $\#\mathbb{F}$ is even, then $P = 1$.
b) If $\#\mathbb{F}$ is odd, then P is the unique element of order 2 in $\mathbb{F}^\times$, namely -1.
So for any odd prime p, the second case holds for the field $\mathbb{Z}/p\mathbb{Z}$: Wilson's Theorem.

As we mentioned above, Wilson's Theorem construed as a statement about the product of all residue classes from 1 up to $n-1$ modulo n holds exactly when n is prime. On the other hand, for composite n we may still apply Theorem B.27 to the finite commutative group $U(n) = (\mathbb{Z}/n\mathbb{Z})^\times$.

THEOREM B.30. *(Gauss) Let $n > 2$ be an integer, and let $U(n)$ be the multiplicative group of units of the finite ring $\mathbb{Z}/n\mathbb{Z}$. Put $P = \prod_{x \in U(n)} x$. Then:*
a) We always have $P = \pm 1 \pmod n$.
b) More precisely: $P = -1 \pmod n$ if and only if n is 4, an odd prime power p^b, or twice an odd prime power $2p^b$.

PROOF. a) For $n > 2$, $-1 \pmod n$ is an element of order 2 in $U(n)$; applying Theorem B.27 to $G = U(n)$, we get $P = \pm 1$: further, $P = -1$ if -1 is the only order 2 element of $U(n)$, and $P = 1$ if there is a further element of order 2.
b) Let $n = 2^a p_1^{b_1} \cdots p_r^{b_r}$ with $p_1 < \ldots < p_r$ odd prime numbers, $a \in \mathbb{N}$ and $b_1, \ldots, b_r \in \mathbb{Z}^+$. By the Chinese Remainder Theorem,

$$U(n) \cong U(2^a) \times \prod_{i=1}^r U(p_i^{b_i}).$$

Case 1: $r = 0$, $a = 2$. Then $U(4)$ is cyclic of order 2, so $d_2(U(4)) = 1$ and $P = -1$.
Case 2: $a \geq 3$. Then $U(2^a) \subset U(n)$, and $U(n) = U(2^a) \cong Z_{2^{a-2}} \times Z_2$, so by Lemma B.28, $d_2(U(2^a)) = 4$. Thus $d_2(U(n)) \geq 4$ and $P = 1$.
Case 3: $a \leq 1$ and $r = 1$, i.e., $n = p^b$ or $n = 2p^b$ for an odd prime power p^b. The groups $U(1)$ and $U(2)$ are trivial, so $U(n) \cong U(p^n)$. By Theorem 22, $U(p^n)$ is cyclic

[7] We are now talking about multiplicative groups rather than additive groups. It makes no mathematical difference, of course, but the reader may wish to pause to reorient to the new notation.

of even order, so $d_2(U(n)) = 1$ and $P = -1$.
Case 4: Suppose $r \geq 2$. Then $U(p^{b_1}) \times U(p^{b_2}) \subset U(n)$, so

$$d_2(U(n)) \geq d_2(U(p^{b_1}) \times U(p^{b_2})) = d_2(U(p^{b_1})) \times d_2(U(p^{b_2})) = 4.$$

Thus $P = 1$. □

After this section was first written, I found a closely related paper of the early American group theorist George Abram Miller [**Mi03**]. In particular Miller proves Theorem B.27 (with a very similar proof) and applies it to prove Theorem B.30. That this result was first stated and proved by Gauss is not mentioned in Miller's paper, but its title suggests that he may have been aware of this.

APPENDIX C

More on Polynomials

1. Polynomial Rings

Let k be a field, and consider the univariate polynomial ring $k[t]$.

THEOREM C.1. *The ring $k[t]$ is a PID and hence a UFD.*

PROOF. In fact the argument is very close to the one we used to show that $\mathbb{Z}$ is a PID. Namely, let I be an ideal of $k[t]$: we may assume that $I \neq 0$. Let b be a nonzero element of I of minimal degree: we claim that $I = \langle a \rangle$. Indeed, let a be any element of I. By polynomial division (this is the key!), there are $q, r \in k[t]$ such that $a = qb + r$ and $\deg r < \deg b$. Since $a, b \in I$, $r = a - qb \in I$. Since $\deg r < \deg b$ and b has minimal degree among nonzero elements of I, we must have $r = 0$, and thus $a = qb$ and $a \in \langle b \rangle$. Thus $k[t]$ is a PID and hence also a UFD. □

Polynomial differentiation: When $k = \mathbb{R}$, every polyomial $f \in \mathbb{R}[t]$ can be viewed as a differentiable function $f : \mathbb{R} \to \mathbb{R}$, and indeed the derivative f' is again a polynomial. Although the derivative is defined by a limiting process, when restricted to polynomials it is characterized by the following two properties:

(P1): $f \mapsto f'$ is an $\mathbb{R}$-linear map: for all $\alpha, \beta \in \mathbb{R}$ and all polynomials f, g,

$$(\alpha f + \beta g)' = \alpha f' + \beta g'.$$

(P2): $1' = 0$, and for all $n \in \mathbb{Z}^+$, $(t^n)' = nt^{n-1}$.

Indeed, the set $\{1, t^n \mid n \in \mathbb{Z}^+\}$ is a basis for the $\mathbb{R}$-vector space $\mathbb{R}[t]$, and since differentiation is linear, it is entirely determined by its action on this basis.

Now let k be any field. It is still true that $\{1, t^n \mid n \in \mathbb{Z}^+\}$ is a k-basis for $k[t]$, so there is a unique k-linear endomorphism of $k[t]$ defined by $1' = 0$ and $(t^n)' = nt^{n-1}$. We continue to call the operator $f \mapsto f'$ **differentiation** and refer to f' as the **derivative** of f, despite the fact that there are no limits here: it is purely algebraic.

Exercise: Show that for any field k, polynomial differentiation satisfies the **product rule**: for all $f, g \in k[t]$, $(fg)' = f'g + fg'$.

Exercise: Compute the kernel of differentiation as a linear endomorphism of $k[t]$. In more concrete terms, find all polynomials $f \in k[t]$ such that $f' = 0$. (Hint: the answer strongly depends on the characteristic of k.)

We say a polynomial $f \in k[t]$ is **separable** if $\gcd(f, f') = 1$.

PROPOSITION C.2. *A separable polynomial is squarefree.*

PROOF. By contraposition: suppose $f = gh^2$ for a polynomial h of positive degree. Then $f' = (gh^2)' = g'h^2 + g(2hh') = h(g'h + 2gh')$. It follows that h is a common divisor of f and f', so f is not separable. □

Exercise: Let k be a field of characteristic $p > 0$, and let $K = k(x)$ be the field of rational functions with coefficients in k. Consider the polynomial $f = t^p - x \in K[t]$.
a) Show that f is squarefree.
b) Show that f is not separable.

As the previous exercise shows, the converse of Proposition C.2 is not generally valid: a squarefree polynomial need not be separable. Of course the counterexample took place over a rather exotic field. In fact for most of the fields one meets in undergraduate mathematics (and, in particular, in this text) it turns out that all squarefree polynomials are separable. Technically speaking, this holds for polynomials over a field k iff k is **perfect**. The class of perfect fields includes all fields of characteristic zero and all finite fields.

2. Finite Fields

PROPOSITION C.3. *Let $\mathbb{F}$ be a finite field. Then $\#\mathbb{F} = p^a$ for some prime number p and some positive integer a.*

PROOF. Since $\mathbb{F}$ is finite, the elements, 1, $1+1$, $1+1+1, \ldots, 1+\cdots+1$ cannot all be distinct. Thus the characteristic of $\mathbb{F}$ must be positive and then necessarily a prime number p. That is, the subfield of $\mathbb{F}$ generated by 1 may be identified with $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. Then $\mathbb{F}$ is a finite-dimensional vector space over $\mathbb{F}_p$. Let $e_1, \ldots, e_a$ be a basis for $\mathbb{F}$ over $\mathbb{F}_p$. Then every element of $\mathbb{F}$ has a unique expression of the form $\alpha_1 e_1 + \ldots + \alpha_a e_a$ with $\alpha_1, \ldots, \alpha_a \in \mathbb{F}_p$, and it follows that $\#\mathbb{F} = p^a$. □

Exercise: Let p be a prime number, and let R be a commutative ring with $\#R = p$. Show that $R \cong \mathbb{Z}/p\mathbb{Z}$.

The fundamental result on finite fields is the following one.

THEOREM C.4. *Let p be a prime number, a a positive integer, and put $q = p^a$.*
a) There is a finite field $\mathbb{F}$ with $\#\mathbb{F} = q$.
b) Any two finite fields with the same cardinality are isomorphic.

In view of Theorem C.4, for any prime power q we may speak of "the finite field" $\mathbb{F}_q$ of order q.[1]

The proof of Theorem C.4 is relatively easy if one has at hand a basic concept in field theory, that of a **splitting field** for a polynomial f. Then the existence of $\mathbb{F}_{p^a}$ follows from the existence of a splitting field for the polynomial $f(t) = t^{p^a} - t$ over the field $\mathbb{F}_p$, and the fact that any two finite fields of the same order are isomorphic follows from the uniqueness of splitting fields. Here we will give a more concrete proof of Theorem C.4a). In fact we will do better: given any finite field $\mathbb{F}$, we will give a formula for the number of degree a monic irreducible polynomials with coefficients in $\mathbb{F}$. This argument is done in two steps. The first step is a piece of pure field theory; in the second step we apply Möbius Inversion.

[1]To be honest it is slightly sloppy to speak in this way, since $\mathbb{F}_q$ is only unique *up to isomorphism*. But this sloppiness is very standard, and, especially for our purposes, absolutely harmless.

LEMMA C.5. *Let $a, b, x \in \mathbb{Z}^+$ with $x > 1$. The following are equivalent:*
(i) $a \mid b$.
(ii) $x^a - 1 \mid x^b - 1$.

PROOF. Write $b = qa + r$ with $0 \leq r < a$.
(i) $\implies$ (ii): If $a \mid b$, then $r = 0$ and

$$x^b - 1 = (x^a)^q - 1 = (x^a - 1)(1 + x^a + \ldots + (x^a)^{q-1}).$$

(ii) $\implies$ (i): We have

$$x^b - 1 = x^b - x^r + x^r - 1 = x^r(x^{aq} - 1) + x^r - 1.$$

By (i) $\implies$ (ii) we know that $x^a - 1 \mid x^{aq} - 1$. By assumption $x^a - 1 \mid x^b - 1$, so we conclude $x^a - 1 \mid x^r - 1$. Since $r < a$, if $r > 0$ we'd have $x^a - 1 > x^r - 1 > 0$, contradiction. So $r = 0$ and $a \mid b$. □

THEOREM C.6. *Let $\mathbb{F}$ be a finite field of order q, let $a \in \mathbb{Z}^+$, and consider $f = t^{q^a} - t \in \mathbb{F}[t]$. Then f is squarefree, and its factors are precisely the monic irreducible polynomials $P \in \mathbb{F}[t]$ of degree dividing a.*

PROOF. Step 1: We have $f' = q^a t^{q^a - 1} - 1 = -1$, since $\mathbb{F}$ has characteristic p. By Proposition C.2, f is squarefree. Since f is monic, it is therefore a product of distinct monic irreducible polynomials, and our task is now to show that a monic irreducible polynomial P divides $f = t^{q^a} - t$ iff $\deg P \mid a$.
Step 2: For an irreducible polynomial P, put $\mathbb{F}_P = \mathbb{F}[t]/(P)$, a field extension of degree $d = \deg P$. Then $P \mid f \iff t^{q^a} - t = 0 \in \mathbb{F}_P$. Since t generates $\mathbb{F}_P$ over $\mathbb{F}$, if $t^{q^a} = t$, then every element $x \in \mathbb{F}_P$ satisfies $x^{q^a} = x$, or equivalently every nonzero element of $\mathbb{F}_P$ has order dividing $q^a - 1$. Since $\mathbb{F}_p^\times$ is cyclic of order $q^d - 1$, this holds iff $q^d - 1 \mid q^a - 1$ iff (by Lemma C.5) $d \mid a$. □

THEOREM C.7. *Let $\mathbb{F}$ be a finite field of order q, and let $n \in \mathbb{Z}^+$. The number of monic irreducible polynomials of degree n with coefficients in $\mathbb{F}_q$ is*

$$I(\mathbb{F}, n) = \frac{1}{n} \sum_{d \mid n} q^d \mu\left(\frac{n}{d}\right). \tag{67}$$

PROOF. For any $d \in \mathbb{Z}^+$, let $I(\mathbb{F}, d)$ be the number of monic irreducible degree d polynomials with $\mathbb{F}$-coefficients. Let $n \in \mathbb{Z}^+$. Then since $t^{q^n} - t$ is the squarefree product of all monic irreducible polynomials of degrees $d \mid n$, equating degrees gives

$$\sum_{d \mid n} d I(\mathbb{F}, d) = q^n.$$

Applying Möbius Inversion, we get

$$n I(\mathbb{F}, n) = \sum_{d \mid n} q^d \mu\left(\frac{n}{d}\right).$$

Dividing both sides by n we get (67). □

COROLLARY C.8. *a) For any finite field $\mathbb{F}$ and any $n \in \mathbb{Z}^+$, there is at least one irreducible polynomial of degree n with $\mathbb{F}$-coefficients.*
b) For every prime power $q = p^a$, there is a finite field $\mathbb{F}$ of order q.

PROOF. a) Theorem C.7 gives us an expression for the number of monic irreducible polynomials of degree n with $\mathbb{F}$-coefficients. By making some very crude estimates we can quickly see that this quantity is always positive. Indeed:

$$I(\mathbb{F},n)=\frac{1}{n}\sum_{d|n}q^d\mu\left(\frac{n}{d}\right)\geq\frac{1}{n}(q^n-(q^{n-1}+\ldots+q+1))$$

$$=\frac{1}{n}\left(q^n-\frac{q^n-1}{q-1}\right)\geq\frac{1}{n}\left(q^n-(q^n-1)\right)=\frac{1}{n}>0.$$

b) By part a), there is a degree a irreducible polynomial f with $\mathbb{F}_p$-coefficients. Then $\mathbb{F}_p[t]/(f)$ is a finite field of order p^a. □

EXERCISE C.1. Try to extract from (67) more realistic estimates on the size of $I(\mathbb{F},n)$.

Bibliography

[A1] Apostol, Tom M. Introduction to analytic number theory. Undergraduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, 1976.

[A2] Apostol, Tom M. Modular functions and Dirichlet series in number theory. Second edition. Graduate Texts in Mathematics, 41. Springer-Verlag, New York, 1990.

[AD93] N. Alon and M. Dubiner, *Zero-sum sets of prescribed size*. Combinatorics, Paul Erdős is eighty, Vol. 1, 33–50, Bolyai Soc. Math. Stud., Janos Bolyai Math. Soc., Budapest, 1993.

[Al99] N. Alon, *Combinatorial Nullstellensatz*. Recent trends in combinatorics (Mátraháza, 1995). Combin. Probab. Comput. 8 (1999), 7–29.

[An57] N.C. Ankeny, *Sums of three squares*. Proc. Amer. Math. Soc. 8 (1957), 316-319.

[AT92] N. Alon and M. Tarsi *Colorings and orientations of graphs*. Combinatorica 12 (1992), 125–134.

[Au12] L. Aubry, Sphinx-Œdipe 7 (1912), 81–84.

[Ax64] J. Ax, *Zeroes of polynomials over finite fields*. Amer. J. Math. 86 (1964), 255–261.

[BR89] C. Bailey and R.B. Richter, *Sum zero (mod n), size n subsets of integers*. Amer. Math. Monthly 96 (1989), 2400–242.

[Ba11] M. Baker, *Zolotarev's magical proof of the law of quadratic reciprocity*. 2011

[BG75] E.A. Bender and J.R. Goldman, *On the applications of Mb̈ius inversion in combinatorial analysis*. Amer. Math. Monthly 82 (1975), 789-803.

[BC12] A. Brunyate and P.L. Clark, *Extending the Zolotarev-Frobenius approach to quadratic reciprocity*. Ramanujan J. 37 (2015), 25-50.

[Bh00] M. Bhargava, *On the Conway-Schneeberger fifteen theorem*. Quadratic forms and their applications (Dublin, 1999), 27–37, Contemp. Math., 272, Amer. Math. Soc., Providence, RI, 2000.

[BHxx] M. Bhargava and J.P. Hanke, *Universal quadratic forms and the 290-theorem*, to appear in Invent. Math.

[Bl14] H.F. Blichfeldt, *A new principle in the geometry of numbers, with some applications*. Trans. Amer. Math. Soc. 15 (1914), 227–235.

[BR51] A. Brauer and R.L. Reynolds, *On a theorem of Aubry-Thue*. Canadian J. Math. 3 (1951), 367–374.

[BTxx] D.G. Best and T.S. Trudgian, *Linear relations of zeroes of the zeta-function*. To appear in Mathematics of Computation.

[CaGN] J.W.S. Cassels, *An introduction to the geometry of numbers* [corrected reprint of the 1971 edition]. Classics in Mathematics. Berlin: Springer-Verlag; 1997.

[CaQF] J.W.S. Cassels, *Rational quadratic forms*. London: Academic Press; 1978.

[CE59] E.D. Cashwell and C.J. Everett, *The ring of number-theoretic functions*. Pacific J. Math. 9 (1959) 975–985.

[Ch36] C. Chevalley, *Démonstration d'une hypothèse de M. Artin*. Abh. Math. Sem. Univ. Hamburg 11 (1936), 73–75.

[CJ14] P.L. Clark and W.C. Jagy, *Euclidean quadratic forms and ADC forms II: integral forms*. Acta Arith. 164 (2014), 265–308.

[Cl94] D.A. Clark, *A quadratic field which is Euclidean but not norm-Euclidean*. Manuscripta Math. 83 (1994), no. 3-4, 327–330.

[Cl09] P.L. Clark, *Elliptic Dedekind domains revisited*. Enseignement Math. 55 (2009), 213–225.

[Cl12] P.L. Clark, *Euclidean Quadratic Forms and ADC-forms I*. Acta Arithmetica 154 (2012), 137–159.

[Cl19] P.L. Clark, *Rabinowitsch times six*, `alpha.math.uga.edu/~pete/Rabinowitsch.pdf`.

[CM98] T. Cochrane and P. Mitchell, *Small solutions of the Legendre equation*. J. Number Theory 70 (1998), 62–66.

[Cox] D.A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory and complex multiplication*. New York: John Wiley & Sons, Inc.; 1989.

[Coh73] P.M. Cohn, *Unique factorization domains*. Amer. Math. Monthly 80 (1973), 1–18.

[Conr-A] K. Conrad, *Two applications of unique factorization*. `http://www.math.uconn.edu/~kconrad/blurbs/ringtheory/ufdapp.pdf`

[Conr-B] K. Conrad, *Examples of Mordell's Equation*. `http://www.math.uconn.edu/~kconrad/blurbs/gradnumthy/mordelleqn1.pdf`

[Con97] J.H. Conway, *The sensual (quadratic) form*. With the assistance of Francis Y. C. Fung. Carus Mathematical Monographs, 26. Mathematical Association of America, Washington, DC, 1997.

[Con00] J.H. Conway, *Universal quadratic forms and the fifteen theorem*. Quadratic forms and their applications (Dublin, 1999), 23–26, Contemp. Math., 272, Amer. Math. Soc., Providence, RI, 2000.

[CS07] W. Cao and Q. Sun, *Improvements upon the Chevalley-Warning-Ax-Katz-type estimates*. J. Number Theory 122 (2007), 135–141.

[Dic27] L.E. Dickson, *Integers represented by positive ternary quadratic forms*. Bull. Amer. Math. Soc. 33 (1927), 63–70.

[DH05] W. Duke and K. Hopkins, *Quadratic reciprocity in a finite group*. Amer. Math. Monthly 112 (2005), no. 3, 251–256.

[DV09] S. Dasgupta and J. Voight, *Heegner points and Sylvester's conjecture*. Arithmetic geometry, 91–102, Clay Math. Proc., 8, Amer. Math. Soc., Providence, RI, 2009.

[Eh55] E. Ehrhrart, *Une généralisation du théorème de Minkowski*. C. R. Acad. Sci. Paris 240 (1955), 483–485.

[Euc] Euclid, *The thirteen books of Euclid's Elements translated from the text of Heiberg. Vol. I: Introduction and Books I, II. Vol. II: Books III–IX. Vol. III: Books X–XIII and Appendix*. Translated with introduction and commentary by Thomas L. Heath. 2nd ed. Dover Publications, Inc., New York, 1956.

[EGZ61] P. Erdős, A. Ginzburg and A. Ziv, *Theorem in the additive number theory*. Bull. Research Council Israel 10F (1961), 41–43.

[F] D.E. Flath, *Introduction to Number Theory*. Wiley-Interscience Publications, 1989.

[Fu55] H. Furstenberg, *On the infinitude of primes*. Amer. Math. Monthly 62 (1955), 353.

[Go59] S.W. Golomb, *A Connected Topology for the Integers*. Amer. Math. Monthly 66 (1959), 663–665.

[GC68] I.J. Good and R.F. Churchhouse, *The Riemann hypothesis and pseudorandom features of the Möbius sequence*. Math. Comp. 22 (1968), 857–861.

[GM84] R. Gupta and M.R. Murty, *A remark on Artin's conjecture*. Invent. Math. 78 (1984), 127–130.

[GPZ98] J. Gebel, A. Pethö and H.G. Zimmer, *On Mordell's equation*. Compositio Math. 110 (1998), 335–367.

[GT08] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*. Ann. of Math. (2) 167 (2008), 481–547.

[Hag11] T. Hagedorn, *Primes of the form $x^2 + ny^2$ and the geometry of (convenient) numbers*, preprint.

[Ham68] J. Hammer, *On some analogies to a theorem of Blichfeldt in the geomety of numbers*. Amer. Math. Monthly 75 (1968), 157–160.

[Han04] J.P. Hanke, *Local densities and explicit bounds for representability by a quadratric form*. Duke Math. J. 124 (2004), 351–388.

[Har73] H. Harborth, *Ein Extremalproblem für Gitterpunkte*. Collection of articles dedicated to Helmut Hasse on his seventy-fifth birthday. J. Reine Angew. Math. 262/263 (1973), 356–360.

[Has28] H. Hasse, *Über eindeutige Zerlegung in Primelemente oder in Primhauptideale in Integrit?tsbereichen*. J. reine Angew. Math. 159 (1928), 3–12.

[HB86] D.R. Heath-Brown, *Artin's conjecture for primitive roots*. Quart. J. Math. Oxford Ser. (2) 37 (1986), 27–38.

[Ho50] L. Holzer, *Minimal solutions of Diophantine equations.* Canadian J. Math. 2 (1950), 238–244.

[Hu00] M.N. Huxley, *Integer points in plane regions and exponential sums.* Number theory, 157–166, Trends Math., Birkhäuser, Basel, 2000.

[HW] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers.* Fifth edition. The Clarendon Press, Oxford University Press, New York, 1979.

[IR] K. Ireland and M. Rosen, *A classical introduction to modern number theory.* 2nd ed. New York: Springer-Verlag; 1990.

[Ka71] N.M. Katz, *On a theorem of Ax.* Amer. J. Math. 93 (1971), 485–499.

[Ká05] G. Károlyi, *Cauchy-Davenport theorem in group extensions.* Enseign. Math. (2) 51 (2005), 239–254.

[Ke83] A. Kemnitz, *On a lattice point problem.* Ars Combin. 16 (1983), B, 151–160.

[KvdL04] T. Kotnik and J. van de Lune, *On the order of the Mertens function.* Experiment. Math. 13 (2004), 473–481.

[Le02] H.W. Lenstra, *Solving the Pell equation.* Notices Amer. Math. Soc. 49 (2002), 182–192.

[Li33] F.A Lindemann, *The Unique Factorization of a Positive Integer.* Quart. J. Math. 4, 319-320, 1933.

[LF73] H. London and R. Finkelstein, *On Mordell's equation* $y^2 - k = x^3$. Bowling Green State University, Bowling Green, Ohio, 1973.

[Le96] M. Lerch, *Sur un théorème de Zolotarev.* Bull. Intern. de l'Acad. François Joseph 3 (1896), 34–37.

[Mar71] C.F. Martin, *Unique factorization of arithmetic functions.* Aequationes Math. 7 (1971), 211.

[Me09] I.D. Mercer, *On Furstenberg's proof of the infinitude of primes.* Amer. Math. Monthly 116 (2009), 355–356.

[Mi03] G.A. Miller, *A new proof of the generalized Wilson's theorem.* Ann. of Math. (2) 4 (1903), 188–190.

[Mi10] H. Minkowski, *Geometrie der Zahlen.* Zweite Lieferung. 1910. B. G. Teubner, Leipzig und Berlin.

[Mi27] H. Minkowski, *Diophantische Approximationen* 1927, Leipzig-Berlin

[M] L.J. Mordell, *Diophantine equations.* Pure and Applied Mathematics, Vol. 30 Academic Press, London-New York, 1969.

[Mo34] L.J. Mordell, *On some arithmetical results in the geometry of numbers.* Compositio Math. 1 (1935), 248–253.

[Mo69] L.J. Mordell, *On the magnitude of the integer solutions of the equation* $ax^2+by^2+cz^2 = 0$. J. Number Theory 1 (1969), 1–3.

[Mo79] P. Morton, *A generalization of Zolotarev's theorem.* Amer. Math. Monthly 86 (1979), 374–375.

[MT06] M.R. Murty and N. Thain, *Prime numbers in certain arithmetic progressions.* Funct. Approx. Comment. Math. 35 (2006), 249–259.

[MT07] M.R. Murty and N. Thain, *Pick's theorem via Minkowski's theorem.* Amer. Math. Monthly 114 (2007), 732–736.

[Nag] T. Nagell, *Introduction to number theory.* 2nd ed. New York: Chelsea Publishing Company; 1964.

[Nar] W. Narkiewicz. *Elementary and analytic theory of algebraic numbers.* Third edition. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004.

[NZM] I. Niven, H.S. Zuckerman and H.L. Montgomery. *An introduction to the theory of numbers.* 5th ed. New York: John Wiley & Sons, Inc.; 1991.

[Nu10] L.M. Nunley, *Geometry of Numbers Approach to Small Solutions of the Extended Legendre Equation.* UGA Master's thesis, 2010.

[Ol76] J.E. Olson, *On a combinatorial problem of Erdős, Ginzburg, and Ziv.* J. Number Theory 8 (1976), 52–57.

[OlR85] A.M. Odlyzko and H.J.J. te Riele *Disproof of the Mertens conjecture.* J. Reine Angew. Math. 357 (1985), 138–160.

[Pi99] G. Pick, Geometrisches zur Zahlenlehre, *Zeit. Vereines 'Lotos' 19* (1899), 311–319.

[Pi87] J. Pintz. *An effective disproof of the Mertens conjecture.* Astérisque 147–148 (1987), 325–333.

[Ra17] Ramanujan, *On the expression of a number in the form* $ax^2 + by^2 + cz^2 + du^2$, Proceedings of the Cambridge Philosophical Society 19(1917), 11–21. See also `http://en.wikisource.org/wiki/Proceedings_of_the_Cambridge_Philosophical_Society`

[Re07] C. Reiher, *On Kemnitz' conjecture concerning lattice-points in the plane*. Ramanujan J. 13 (2007), 333–337.

[Ro63] K. Rogers, *Classroom notes: Unique Factorization*. Amer. Math. Monthly 70 (1963), 547–548.

[RS62] J.B. Rosser and L. Schoenfeld, *Approximate formulas for some functions of prime numbers*. Illinois J. Math. 6 (1962), 64–94.

[Sa68] P. Samuel, *Unique factorization*. Amer. Math. Monthly 75 (1968), 945–952.

[SatR14] Yannick Saouter and H. te Riele, *Improved results on the Mertens conjecture*. Math. Comp. 83 (2014), 421–433.

[Se73] J.-P. Serre, *A course in arithmetic*. Translated from the French. Graduate Texts in Mathematics, No. 7. Springer-Verlag, New York-Heidelberg, 1973.

[SiGN] C.L. Siegel, *Lectures on the geometry of numbers*. Berlin: Springer-Verlag; 1989.

[St] R.P. Stanley, *Enumerative combinatorics. Vol. 1.* With a foreword by Gian-Carlo Rota. Corrected reprint of the 1986 original. Cambridge Studies in Advanced Mathematics, 49. Cambridge University Press, Cambridge, 1997.

[St67] H.M. Stark, *A complete determination of the complex quadratic fields of class-number one*. Michigan Math. J. 14 1967 1–27.

[St-ANT] P. Stevenhagen, *Number Rings*. Course notes available at `http://websites.math.leidenuniv.nl/algebra/ant.pdf`.

[Su95] D.B. Surowski, *The Uniqueness Aspect of the Fundamental Theorem of Finite Abelian Groups*. Amer. Math. Monthly, 102 (1995), 162–163.

[Su99] B. Sury, *The Chevalley-Warning theorem and a combinatorial question on finite groups*. Proc. Amer. Math. Soc. 127 (1999), 951–953.

[Sy84] J. J. Sylvester. *Mathematical Questions, with their solutions*, Educational Times 41 (1884), 21.

[T] G. Tenenbaum, *Introduction to analytic and probabilistic number theory*. Translated from the second French edition (1995) by C. B. Thomas. Cambridge Studies in Advanced Mathematics, 46. Cambridge University Press, Cambridge, 1995.

[Tr88] H.F. Trotter, *An overlooked example of nonunique factorization*. Amer. Math. Monthly 95 (1988), no. 4, 339–342.

[Vi27] I.M. Vinogradov, *On a general theorem concerning the distribution of the residues and non-residues of powers*. Trans. Amer. Math. Soc. 29 (1927), 209–217.

[Wa36] E. Warning, *Bemerkung zur vorstehenden Arbeit von Herrn Chevalley*. Abh. Math. Sem. Hamburg 11 (1936), 76–83.

[W] A. Weil, *Number theory. An approach through history from Hammurapi to Legendre*. Reprint of the 1984 edition. Modern Birkhäuser Classics, Boston, MA, 2007.

[Wh12] J.P. Wheeler, *The Cauchy-Davenport Theorem for Finite Groups*. Preprint available at `http://arxiv.org/abs/1202.1816`.

[Wó72] J. Wójcik, *On sums of three squares*. Colloq. Math. 24 (1971/72), 117–119.

[Z34] E. Zermelo, *Elementare Betrachtungen zur Theorie der Primzahlen*. Nachr. Gesellsch. Wissensch. Göttingen 1 (1934), 43–46.

[Zo72] G. Zolotarev, *Nouvelle démonstration de la loi de réciprocité de Legendre*. Nouvelles Ann. Math. (2) 11 (1872), 354–362.