

Approximation by Rationals

*Shixuan Cao, Ben Mackay, Liam
Power Albalat, Jingren Zhou*

Year 4 Project
School of Mathematics
University of Edinburgh
April 21, 2026

Abstract

We study rational approximation of real numbers. In particular, we show that every irrational number is approximable to order 2. We then consider stronger approximation properties for algebraic numbers. In this context we present two fundamental results preceding Roth's theorem, namely Liouville's theorem and Thue's theorem, together with some of their applications. More specifically, we explore the link between rational approximation and the scrutiny of binary homogeneous polynomials and Pell equations. Finally, we discuss Dirichlet's Unit Theorem, which plays a central role in describing the structure of solutions to Pell equations.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text. Furthermore, Generative AI has only been used to assist with L^AT_EX formatting. The tool used was ChatGPT.

(*Shixuan Cao, Ben Mackay, Liam Power Albalat, Jingren Zhou*)

*We dedicate this report to Professor James Wright, and
extend our gratitude for all of his support.*

Contents

Abstract	ii
Contents	v
1 Introduction	1
2 Foundations of Diophantine Approximation	3
2.1 Approximating Rationals	3
2.2 Approximating Irrationals	5
2.3 Continued Fractions	6
3 Liouville’s Theorem	10
3.1 Preliminaries for Liouville’s Theorem	10
3.2 Liouville’s Theorem	12
3.3 Liouville Numbers	13
4 Application: Binary Homogeneous Polynomials	14
4.1 Overview and the Role of Liouville’s Theorem	14
4.2 Definitions and Basic Properties	15
4.3 Minimal Polynomials and Multiplicity	16
4.4 Analysis of Binary Forms	17
4.5 Constructing Rational Approximations	18
4.6 Two Exceptional Families	20
5 Thue’s Theorem	23
5.1 Preliminaries for Thue’s Theorem	23
5.2 Bounding the Height of the Polynomial	26
5.2.1 Preliminaries: Content and Gauss’s Lemma	26
5.2.2 Lower Bounds for Height	27
5.3 Thue’s Theorem	31
6 Pell Equations and Dirichlet’s Unit Theorem	36
6.1 Dirichlet’s Unit Theorem	43
6.2 Proof of Dirichlet’s Unit Theorem	47
7 Conclusion	51
Bibliography	52

Chapter 1

Introduction

Given a real number ξ , how well can it be approximated by rationals? Upon first inspection, the answer to this question appears trivial. Indeed, for a given $\varepsilon > 0$, the density of $\mathbb{Q}$ in $\mathbb{R}$ guarantees the existence of some rational p/q satisfying

$$\left| \xi - \frac{p}{q} \right| < \varepsilon.$$

However, an issue with this observation is that it does not distinguish between rationals that are merely *close* to ξ and those which provide genuinely *good approximations* of ξ . To illustrate this difference, consider the problem of approximating π by rationals. The rational number $355/113$ is closer to π than the truncation of π to six decimal places, but not as close as the truncation of π to seven decimal places. To then dismiss $355/113$ in favour of the latter would overlook the remarkable accuracy that $355/113$ achieves relative to its denominator. Hence, when considering prospective approximations, it may be desirable to have some equitable way of comparing them. Precisely, we want to account for the size of the denominator; rational numbers with larger denominators should be required to achieve correspondingly smaller errors. We could do this by considering the set of rationals satisfying the following constrained inequality:

$$\left| \xi - \frac{p}{q} \right| < \frac{1}{q^\kappa} \quad \text{where } \kappa > 0. \tag{1.1}$$

Setting $\kappa = 2$, for example, we find that $355/113$ satisfies the inequality with $\xi = \pi$, but the aforementioned truncations do not. This alludes to our previous discussion. For a given real number ξ , we may then ask how the number of rationals satisfying (1.1) varies as κ grows, and whether or not the characteristics of ξ have any influence on this relationship. By investigating these questions, we can develop a notion of approximability of a real number by rationals to a certain *order*. With this notion in place, we can then compare how well different real numbers can be approximated by rationals. This ultimately forms the basis of Diophantine approximation. We shall also explore the underlying structure of approximations with particularly small errors relative to their denominator, such as $355/113$, which will lead us to the theory of continued fractions.

Once we have laid out the foundations of Diophantine approximation, we will investigate the approximability of algebraic irrational numbers. This is a fascinating piece of mathematical history. Over time, mathematicians proved progressively stronger results

concerning how well algebraic irrational numbers can be approximated by rational numbers. Many of these results had far-reaching consequences. Liouville gave the first such result in 1844 by using single variable integer polynomials. Liouville's result provided the first glimpse into the ever-elusive world of transcendental numbers. This was followed by the work of Thue, who used integer polynomials in two variables. Thue's result was intimately connected with his work on Diophantine equations; it allowed him to strictly classify the solution sets to certain families of such equations. We investigate this further in Chapter 4. These results ultimately culminated in Roth's theorem. Rather appropriately, Roth considered integer polynomials with an arbitrary number of variables and he was awarded a fields medal for his result, which was presented to him in our very own McEwan Hall.

We have not yet defined what it means for a real number to be approximable by rationals to a certain order, however, we shall state below the results we discussed. The reader may refer back to them.

Theorem 1.1 (Liouville, 1844 [1, Theorem 1.1]). *Any algebraic number β of degree d is not approximable by rationals to any order greater than d .*

Theorem 1.2 (Thue, 1909 [2, Lecture 25]). *Any algebraic number β of degree $d \geq 2$ is not approximable by rationals to any order greater than $\frac{d}{2} + 1$.*

Theorem 1.3 (Roth, 1955 [1, Theorem 1.2]). *Any algebraic number β of degree $d \geq 2$ is not approximable to any order greater than 2.*

Chapter 2

Foundations of Diophantine Approximation

In this chapter, we develop the basic framework of Diophantine approximation. We begin by examining how well rational and irrational numbers can be approximated by rational numbers¹. This naturally leads to a formal definition of approximability. We then briefly discuss continued fractions, which provide an explicit method for constructing good rational approximations.

2.1 Approximating Rationals

Before turning to irrational numbers, it is instructive to first consider how rational numbers themselves can be approximated by rationals. We begin by examining approximation errors relative to a prescribed scale.

For a given rational ξ , a natural starting point is to ask how many rationals p/q satisfy the following:

$$\left| \xi - \frac{p}{q} \right| \leq \frac{1}{q}.$$

As we shall see, this mild restriction on the error already yields infinitely many rational approximations.

Theorem 2.1. *For any $\xi \in \mathbb{Q}$, there exist infinitely many rationals p/q such that*

$$\left| \xi - \frac{p}{q} \right| \leq \frac{1}{q}.$$

Proof. Let $\xi = \frac{a}{b}$ with $a \in \mathbb{Z}$, $b \in \mathbb{N}$ and $\gcd(a, b) = 1$. By *Bézout's identity*, there exist integers x_0, y_0 such that $ay_0 - bx_0 = 1$. For each integer $r \in \mathbb{Z}$, set $p = x_0 + ra$ and $q = y_0 + rb$. Choose r large enough so that $q = y_0 + rb > 0$. Then

$$\left| \xi - \frac{p}{q} \right| = \left| \frac{a}{b} - \frac{x_0 + ra}{y_0 + rb} \right| = \frac{|ay_0 - bx_0|}{bq} = \frac{1}{bq} \leq \frac{1}{q}.$$

Moreover, notice that $aq - bp = ay_0 - bx_0 = 1$, so $\gcd(p, q) = 1$.

¹When discussing rational numbers p/q , we assume throughout that $q > 0$ and $\gcd(p, q) = 1$.

Finally, suppose that $r_1 \neq r_2$ produce the same rational number. Then

$$\frac{x_0 + r_1 a}{y_0 + r_1 b} = \frac{x_0 + r_2 a}{y_0 + r_2 b}.$$

Cross-multiplying and simplifying gives $(r_1 - r_2)(ay_0 - bx_0) = 0$, which is impossible since $ay_0 - bx_0 = 1$. Hence all these rationals are distinct.

Since r can be chosen arbitrarily large, this produces infinitely many distinct rational numbers p/q satisfying $|\xi - p/q| \leq 1/q$. $\square$

Let us now consider the stronger bound $1/q^2$. This case is naturally handled by a more general result, namely Dirichlet's Approximation Theorem, which guarantees that any real number ξ admits a rational approximation p/q satisfying a bound of the form $|\xi - p/q| < 1/q^2$.

Theorem 2.2 (Dirichlet's Approximation Theorem [3, Section 11.3]). *For any $\xi \in \mathbb{R}$ and any $N \in \mathbb{N}$, there exist integers p and q with $1 \leq q \leq N$ and $\gcd(p, q) = 1$, such that*

$$|q\xi - p| < \frac{1}{N}.$$

Proof. Consider the $N + 1$ fractional parts

$$\{0 \cdot \xi\}, \{1 \cdot \xi\}, \dots, \{N \cdot \xi\} \in [0, 1).$$

If we split the interval $[0, 1)$ into the N subintervals of length $1/N$:

$$\left[0, \frac{1}{N}\right), \left[\frac{1}{N}, \frac{2}{N}\right), \dots, \left[\frac{N-1}{N}, 1\right).$$

By the pigeonhole principle, two of the fractional parts, say $\{m\xi\}$ and $\{n\xi\}$ with $0 \leq m < n \leq N$, must lie in the same subinterval. Hence

$$|\{n\xi\} - \{m\xi\}| < \frac{1}{N}.$$

Since $\{n\xi\} - \{m\xi\} = (n - m)\xi - k$ for some integer k , set $q_0 = n - m$ and $p_0 = k$. Then $1 \leq q_0 \leq N$ and $|q_0\xi - p_0| < 1/N$. Let $d = \gcd(p_0, q_0)$ and write $p_0 = dp$, $q_0 = dq$. Then $\gcd(p, q) = 1$, and

$$|q\xi - p| = \frac{1}{d}|q_0\xi - p_0| < \frac{1}{dN} \leq \frac{1}{N}.$$

Also, $1 \leq q = \frac{q_0}{d} \leq q_0 \leq N$. $\square$

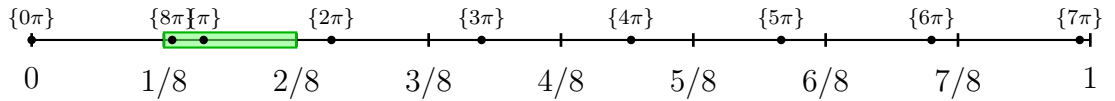


Figure 2.1: Dirichlet's argument applied to $\xi = \pi$ and $N = 8$. We can see that $\{\pi\}$ and $\{8\pi\}$ lie in the same interval.

An immediate consequence of Dirichlet's theorem is the following:

Corollary 2.3 ([3, Section 11.3]). *For any real number ξ , there exists a rational number p/q such that*

$$\left|\xi - \frac{p}{q}\right| < \frac{1}{q^2}. \quad (2.1)$$

Proof. Let $N \in \mathbb{N}$. By Theorem 2.2, there exists a rational number p/q with $1 \leq q \leq N$ such that $|q\xi - p| < 1/N$. Dividing through by q , we obtain

$$\left| \xi - \frac{p}{q} \right| < \frac{1}{Nq} \leq \frac{1}{q^2}.$$

□

Thus, for every real number ξ , there exists at least one rational approximation satisfying (2.1). We now seek to determine whether this stronger inequality holds for infinitely many rationals, as in Theorem 2.1. When ξ is rational, this is not the case:

Lemma 2.4 ([3, Section 11.3]). *For any $\xi \in \mathbb{Q}$ there are only finitely many rationals p/q satisfying the inequality (2.1).*

Proof. Let $\xi = \frac{a}{b}$ where $\gcd(a, b) = 1$. Let p/q be a rational satisfying (2.1). We have

$$\left| \frac{a}{b} - \frac{p}{q} \right| = \left| \frac{aq - bp}{bq} \right| < \frac{1}{q^2} \implies |aq - bp| < \frac{b}{q}.$$

Since $aq - bp$ is an integer, either $aq - bp = 0$ or $|aq - bp| \geq 1$. If $aq - bp = 0$, then $p/q = a/b$, which gives only one approximation. Otherwise,

$$1 \leq |aq - bp| < \frac{b}{q},$$

which implies $q < b$. Hence there are only finitely many possibilities for q . It follows that there are only finitely many rational numbers p/q satisfying (2.1). □

The preceding results suggest that rational approximation should be measured by comparing the error $\left| \xi - \frac{p}{q} \right|$ with a quantity of the form $K(\xi)/q^\kappa$. Here the exponent κ measures the strength of approximation, while the constant $K(\xi) > 0$ allows flexibility in the bound. This motivates the following definition:

Definition 2.5 (Approximability [3]). A real number ξ is said to be *approximable (by rationals) to order $\kappa > 0$* if there exists a constant $K(\xi) > 0$ such that

$$\left| \xi - \frac{p}{q} \right| < \frac{K(\xi)}{q^\kappa}$$

for infinitely many rational numbers p/q in reduced form with $q > 0$.

With this definition in place, our earlier discussion can be summarized as follows: We can deduce from Theorem 2.1 that every rational number is approximable to order $\kappa = 1$. Additionally, Lemma 2.4 suggests that rational numbers may not be approximable to order 2. Although we have not fully established this yet, this statement does turn out to be true.

2.2 Approximating Irrationals

We now turn to irrational numbers. Unlike rational numbers, they are approximable to order 2:

Theorem 2.6 ([3, Theorem 187]). *Every irrational number is approximable to order 2.*

Proof. Suppose that $\xi \in \mathbb{R} \setminus \mathbb{Q}$ is not approximable to order 2. Then, for any constant $K > 0$, there are only finitely many rationals p/q satisfying $|\xi - p/q| < K/q^2$. Consider the case where $K = 1$. Let $\mathcal{S}$ be the set containing all rational numbers satisfying $|\xi - p/q| < 1/q^2$. Then, we can write

$$\mathcal{S} = \left\{ \frac{p_1}{q_1}, \frac{p_2}{q_2}, \dots, \frac{p_s}{q_s} \right\}$$

for some $s \in \mathbb{N}$. Furthermore, we can define

$$r := \min \left\{ \left| \xi - \frac{p_i}{q_i} \right| : \frac{p_i}{q_i} \in \mathcal{S} \right\},$$

and r is strictly positive since ξ is irrational. By the *Archimedean property*, we may choose $Q \in \mathbb{N}$ such that $Qr > 1$, equivalently $r > 1/Q$. Hence, for all $\frac{p_i}{q_i} \in \mathcal{S}$,

$$\left| \xi - \frac{p_i}{q_i} \right| \geq r > \frac{1}{Q}.$$

Applying Dirichlet's Approximation Theorem (Theorem 2.2) to ξ and Q , we obtain integers p and q with $1 \leq q \leq Q$ such that $|q\xi - p| < 1/Q$. Dividing by q gives

$$\left| \xi - \frac{p}{q} \right| < \frac{1}{Qq} \leq \frac{1}{q^2}.$$

Hence, $p/q \in \mathcal{S}$. From the construction above, we have $\left| \xi - \frac{p}{q} \right| < \frac{1}{Q}$. This contradicts minimality of r . Therefore, the set $\mathcal{S}$ cannot be finite. $\square$

Given an irrational ξ , the above result guarantees the existence of infinitely many rationals p/q satisfying $|\xi - p/q| < 1/q^2$. However, it does not provide a method for explicitly constructing these approximations. This leads naturally to continued fractions, which furnish an explicit sequence of rational approximations and, in particular, recover the quadratic bound appearing in Dirichlet's theorem.

2.3 Continued Fractions

In the introduction, we observed that $\frac{355}{113}$ provides a particularly accurate rational approximation of π . The study of continued fractions not only explains where this approximation comes from, but also provides a systematic way to construct others of comparable quality.

Definition 2.7 (Continued Fraction Expansion [3, Section 10.9]). For any real number x , its *simple continued fraction expansion* is defined recursively as follows. Set $x_0 = x$. For each $n \geq 0$, define

$$a_n = \lfloor x_n \rfloor \quad \text{and} \quad x_{n+1} = \frac{1}{x_n - a_n} \quad \text{if } x_n \neq a_n,$$

and terminate the process if $x_n = a_n$ for some n . Thus $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for all $n \geq 1$. In this way one obtains a finite or infinite continued fraction expansion, written

$$x = [a_0; a_1, a_2, \dots].$$

A classical result [3, Chapter X] states that every real number admits a simple continued fraction expansion

$$x = [a_0; a_1, a_2, \dots] := a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots}},$$

which is infinite if and only if x is irrational. We define the n th convergent of x by $\frac{p_n}{q_n} := [a_0; a_1, \dots, a_n]$. In the case where x is irrational, the convergents $\frac{p_n}{q_n}$ converge to x . The following lemma records the basic recursive structure of these convergents:

Lemma 2.8 ([3, Theorems 170–176]). *For each $n \geq 0$, the finite continued fraction $[a_0; a_1, \dots, a_n]$ is a rational number and can be expressed as p_n/q_n , where the integers p_n and q_n are defined by*

$$(p_{-1}, q_{-1}) = (1, 0), \quad (p_0, q_0) = (a_0, 1).$$

For $n \geq 1$, we have

$$p_n = a_n p_{n-1} + p_{n-2}, \quad q_n = a_n q_{n-1} + q_{n-2}$$

and, moreover,

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}.$$

In particular, $q_n > 0$ for all $n \geq 0$, and $q_{n+1} > q_n$ for all $n \geq 1$.

Proof. The result follows by induction on n . The base cases for $n = 0$ and $n = 1$ are immediate from the definition. Assuming the formula holds for $n - 1$ and $n - 2$, writing

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

gives the first two identities.

Similarly, the identity $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$ is obtained by induction, and the full details of the proof are provided in the Supplementary Material, Lemma S.1.3. $\square$

Lemma 2.8 establishes some fundamental properties of continued fractions — namely, that the n th convergent of x is rational. When x is irrational, these convergents provide rational approximations of x in an interesting way. Specifically, successive convergents seem to ‘sandwich’ x , as discussed in the following lemma:

Lemma 2.9 ([3, Theorems 170–176]). *Let $x \in \mathbb{R} \setminus \mathbb{Q}$, and let p_n/q_n and p_{n+1}/q_{n+1} be successive convergents in the continued fraction expansion of x . Then*

$$\frac{p_n}{q_n} < x < \frac{p_{n+1}}{q_{n+1}} \quad \text{or} \quad \frac{p_{n+1}}{q_{n+1}} < x < \frac{p_n}{q_n}.$$

In other words, successive convergents lie on opposite sides of x .

Figure 2.2 illustrates this phenomenon in the irrational case, where the zig-zag curve represents the sequence of convergents p_n/q_n and the dashed line represents x .

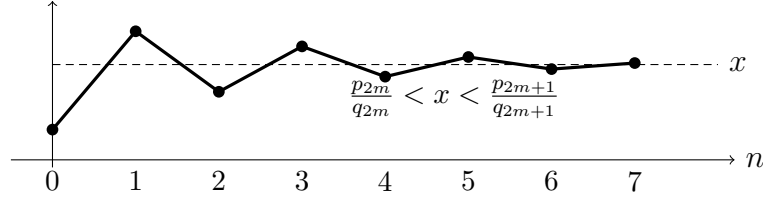


Figure 2.2: Successive convergents alternately lie below and above x .

Proof. Since x is irrational, its simple continued fraction expansion is infinite, so we may write $x = [a_0; a_1, \dots, a_n, x_{n+1}]$, where $x_{n+1} = [a_{n+1}; a_{n+2}, \dots]$. By definition of the continued fraction expansion, we have $x_{n+1} > 0$. Using the recursive structure of the convergents established in Lemma 2.8, we can express x as

$$x = \frac{p_n x_{n+1} + p_{n-1}}{q_n x_{n+1} + q_{n-1}} \implies x - \frac{p_n}{q_n} = \frac{p_n x_{n+1} + p_{n-1}}{q_n x_{n+1} + q_{n-1}} - \frac{p_n}{q_n} = \frac{p_{n-1} q_n - p_n q_{n-1}}{q_n (q_n x_{n+1} + q_{n-1})}.$$

By Lemma 2.8, $p_{n-1} q_n - p_n q_{n-1} = (-1)^n$, and the denominator is strictly positive (since $q_n, q_{n-1} > 0$ and $x_{n+1} > 0$). Hence $\text{sgn}\left(x - \frac{p_n}{q_n}\right) = (-1)^n$. In particular,

$$\left(x - \frac{p_n}{q_n}\right) \left(x - \frac{p_{n+1}}{q_{n+1}}\right) < 0,$$

which shows that x lies strictly between p_n/q_n and p_{n+1}/q_{n+1} . $\square$

We now return to Diophantine approximation. The next result shows that, for irrational x , the convergents p_n/q_n satisfy the quadratic inequality appearing in Dirichlet's theorem, and hence provide an explicit infinite family of such approximations for every irrational number:

Theorem 2.10 ([3, Theorem 171]). *Let x be irrational and let p_n/q_n denote its n th convergent. Then for all $n \geq 1$,*

$$\left|x - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}} < \frac{1}{q_n^2}.$$

Proof. By Lemma 2.8, one can express the difference between two successive convergents as

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

By Lemma 2.9, successive convergents lie on opposite sides of the irrational number x . Hence

$$\left|x - \frac{p_n}{q_n}\right| < \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right|.$$

Finally, since Lemma 2.8 also shows that $q_{n+1} > q_n$ for $n \geq 1$, we have $q_n q_{n+1} > q_n^2$ and hence obtain the desired inequality. $\square$

This explains the exceptional quality of the classical approximation $\frac{355}{113}$ to π , introduced by the Chinese mathematician Zu Chongzhi around CE 470, and known as *Milü*.

The connection between these two constants can be observed via a comparison of their respective continued fraction expansions:

$$\frac{355}{113} = [3; 7, 15, 1] \quad \text{and} \quad \pi = [3; 7, 15, 1, 292, 1, \dots].$$

In particular, $355/113$ is a convergent of the simple continued fraction expansion of π , and the following unusually large next term in the continued fraction, 292, explains why this approximation is so accurate.

The results above apply to any irrational number. To the modern mathematician, the natural next step is to distinguish between algebraic and transcendental irrationals, with the aim of further characterising the real numbers in terms of approximability. Remarkably, the study of rational approximation itself led to this fundamental partition of the irrational numbers. In the following chapters, we shall examine the approximability of algebraic irrational numbers and explore how this endeavour led to the proof of the existence of transcendental numbers.

Chapter 3

Liouville's Theorem

In the previous chapter, we established that all irrational numbers are approximable to at least order 2. This begs the question of whether certain irrationals are approximable to higher order. We shall restrict our attention to the approximability of algebraic irrational numbers.

Liouville was able to provide the first step towards answering this question. In 1844, Liouville proved that any algebraic irrational number of degree d cannot be approximated to any order greater than d . His argument relied on constructing suitable integer polynomials, an idea that later evolved into what is now known as the polynomial method. In recent decades, this method has played a central role in several major breakthroughs. We begin by establishing several preliminary results that are needed for this construction. As a consequence, we can establish the existence of numbers that can be approximated extremely well by rationals. These numbers are now known as Liouville numbers, and were the first explicit examples of transcendental numbers.

3.1 Preliminaries for Liouville's Theorem

For notational convenience, we introduce the following definition.

Definition 3.1. Let $P(x) := \sum_{i=0}^d q_i x^i$ be a polynomial with rational coefficients. The *height* $|P|$ of P is defined by

$$|P| := \max\{|q_i| : i \in \{0, \dots, d\}\}.$$

The following lemma will be particularly useful for the proof of Liouville's Theorem and in the proof of Thue's Theorem:

Lemma 3.2. Let β be an algebraic number of degree d with minimal polynomial $m(x) = \sum_{i=0}^d q_i x^i$. For any integer $n \geq 0$, there exist integers $A_{0,n}, \dots, A_{d-1,n}$ such that

$$q_d^n \beta^n = \sum_{i=0}^{d-1} A_{i,n} \beta^i,$$

and $|A_{i,n}| \leq (2|m|)^n$ for each $i \in \{0, \dots, d-1\}$.

Proof. The proof is somewhat technical and is therefore deferred to the Supplementary Material (Lemma S.1.1). $\square$

We can now use Lemma 3.2 to help us construct a polynomial that will play a key role in the proof of Liouville's Theorem:

Lemma 3.3 ([4]). *Let β be an algebraic number of degree d , and let $\ell \in \mathbb{N}$. Then there exists a non-zero polynomial $P(x) = \sum_{i=0}^{\ell d} b_i x^i \in \mathbb{Z}[x]$ such that*

- (i) $P^{(j)}(\beta) = 0$ for $j = 0, 1, \dots, \ell - 1$;
- (ii) $\deg(P) \leq \ell d$.

Proof. Let $m(x) = \sum_{i=0}^d q_i x^i \in \mathbb{Z}[x]$ be the minimal polynomial of β . Consider the field extension $\mathbb{Q}(\beta)/\mathbb{Q}$, which can be viewed as a $\mathbb{Q}$ -vector space of dimension d with basis $\{1, \beta, \beta^2, \dots, \beta^{d-1}\}$. To satisfy condition (i), we seek a polynomial $P(x)$ of degree D such that

$$\frac{1}{j!} P^{(j)}(\beta) = \sum_{i=j}^D b_i \binom{i}{j} \beta^{i-j} = 0 \quad \text{for } 0 \leq j \leq \ell - 1, \quad (3.1)$$

where we require $\mathbf{b} = (b_0, b_1, \dots, b_D) \in \mathbb{Z}^{D+1}$ and $\mathbf{b} \neq \mathbf{0}$. Let $r \geq 0$. By Lemma 3.2, we have

$$\beta^r = \frac{1}{q_d^r} \sum_{k=0}^{d-1} A_{k,r} \beta^k$$

for some $A_{0,r}, \dots, A_{d-1,r} \in \mathbb{Z}$. Hence, applying this with $r = i - j \geq 0$, we can deduce from (3.1) that

$$\sum_{i=j}^D b_i \binom{i}{j} \left(\frac{1}{q_d^{i-j}} \sum_{k=0}^{d-1} A_{k,i-j} \beta^k \right) = 0 \quad \implies \quad \sum_{k=0}^{d-1} \left(\sum_{i=j}^D b_i \binom{i}{j} \frac{A_{k,i-j}}{q_d^{i-j}} \right) \beta^k = 0.$$

Upon re-indexing, this becomes

$$\sum_{k=0}^{d-1} \left(\sum_{i=0}^{D-j} b_{i+j} \binom{i+j}{j} \frac{A_{k,i}}{q_d^i} \right) \beta^k = 0.$$

Since the set $\{1, \beta, \beta^2, \dots, \beta^{d-1}\}$ is linearly independent over $\mathbb{Q}$, it follows that

$$\sum_{i=0}^{D-j} b_{i+j} \binom{i+j}{j} \frac{A_{k,i}}{q_d^i} = 0 \quad \text{for } k = 0, 1, \dots, d-1.$$

If we multiply each equation through by q_d^D , the above becomes a set of equations over the integers:

$$\sum_{i=0}^{D-j} b_{i+j} \binom{i+j}{j} q_d^{D-i} A_{k,i} = 0 \quad \text{for } k = 0, 1, \dots, d-1.$$

Hence we obtain ℓd homogeneous linear equations with integer coefficients in the $D+1$ unknowns $b_0, \dots, b_D$. Equivalently, we have $L\mathbf{b} = \mathbf{0}$ for some linear map $L : \mathbb{Z}^{D+1} \rightarrow \mathbb{Z}^{\ell d}$. Viewing the same system over $\mathbb{Q}$, if $D+1 > \ell d$, then this homogeneous system has a non-zero solution over $\mathbb{Q}$, and clearing denominators yields a non-zero integer solution. Thus, setting $D = \ell d$, we obtain a non-zero polynomial $P(x) \in \mathbb{Z}[x]$ satisfying $P^{(j)}(\beta) = 0$ for $j = 0, 1, \dots, \ell - 1$, and clearly $\deg(P) \leq \ell d$. $\square$

3.2 Liouville's Theorem

With this polynomial, we can now prove Liouville's theorem.

Theorem 3.4 (Liouville's Theorem [1, Theorem 1.1]). *Any algebraic number β of degree d is not approximable by rationals to any order greater than d .*

Proof. Let β be an algebraic number of degree d . By Lemma 3.3, we can construct a non-zero polynomial $P(x) = \sum_{i=0}^D a_i x^i \in \mathbb{Z}[x]$ with $D \leq \ell d$ satisfying conditions (i) and (ii) for some fixed integer $\ell \geq 1$.

Suppose that β is approximable to order $\kappa > 0$. Then, by definition, there exists a constant $K(\beta) > 0$ such that

$$\left| \beta - \frac{p}{q} \right| < \frac{K(\beta)}{q^\kappa}, \quad (3.2)$$

for infinitely many $\frac{p}{q} \in \mathbb{Q}$. Since there are infinitely many such approximations p/q , their denominators q must be unbounded. Hence we may choose infinitely many approximations p/q such that $q^\kappa > K(\beta)$. Moreover, since P has only finitely many roots, there are still infinitely many such approximations satisfying the additional condition $P\left(\frac{p}{q}\right) \neq 0$.

For any approximation $r := p/q$ satisfying these conditions, we have

$$|P(r)| = \left| \sum_{i=0}^D a_i \left(\frac{p}{q}\right)^i \right| = \frac{\left| \sum_{i=0}^D a_i p^i q^{D-i} \right|}{q^D} \geq \frac{1}{q^D}$$

since $P(r) \neq 0$ and a_i, p, q are integers. This gives a lower bound for $|P(r)|$.

To obtain an upper bound, note that $P(x)$ is ℓ times differentiable on the interval between r and β . Since $q^\kappa > K(\beta)$, it follows from (3.2) that $|\beta - r| < 1$. Hence, by *Taylor's theorem with Lagrange remainder*, there exists some ξ with

$$\beta - 1 < \min\{\beta, r\} \leq \xi \leq \max\{\beta, r\} < \beta + 1$$

such that

$$|P(r)| = \left| \sum_{j=0}^{\ell-1} \frac{P^{(j)}(\beta)}{j!} (r - \beta)^j + \frac{P^{(\ell)}(\xi)}{\ell!} (r - \beta)^\ell \right| = \left| \frac{P^{(\ell)}(\xi)}{\ell!} (r - \beta)^\ell \right|$$

by condition (i). Since polynomials are smooth and $\xi \in [\beta - 1, \beta + 1]$, there exists a constant $C(\beta) \in \mathbb{R}_{>0}$ such that $\left| \frac{P^{(\ell)}(\xi)}{\ell!} \right| < C(\beta)$. It follows that

$$|P(r)| \leq C(\beta) |r - \beta|^\ell < C(\beta) \frac{K(\beta)^\ell}{q^{\kappa\ell}}.$$

Setting $C'(\beta) = C(\beta)K(\beta)^\ell$ and comparing the upper and lower bounds, we obtain

$$\frac{1}{q^D} \leq |P(r)| < C'(\beta) \frac{1}{q^{\kappa\ell}} \implies q^{\kappa\ell - D} < C'(\beta). \quad (3.3)$$

We now show that $\kappa \leq d$. Suppose, for a contradiction, that $\kappa\ell - D > 0$. Since there are infinitely many such rational approximations $r = p/q$, their denominators are unbounded. Hence we can choose r such that $q > C'(\beta)^{\frac{1}{\kappa\ell - D}}$. Since $\kappa\ell - D > 0$, it follows

that

$$q^{\kappa\ell-D} > \left(C'(\beta)^{\frac{1}{\kappa\ell-D}}\right)^{\kappa\ell-D} = C'(\beta),$$

which contradicts (3.3). Hence $\kappa\ell - D \leq 0$, so $\kappa\ell \leq D$. Since $D \leq \ell d$ by condition (ii), we obtain $\kappa\ell \leq \ell d$, and therefore $\kappa \leq d$. $\square$

3.3 Liouville Numbers

The existence of transcendental numbers remained an open question until Liouville used his theorem to construct an explicit class of such numbers. These numbers, now known as Liouville numbers, were the first examples of transcendental numbers and represent a major milestone in the history of mathematics.

Definition 3.5. A real number x is called a *Liouville number* if for all $n \in \mathbb{N}$ there exists $\frac{p}{q} \in \mathbb{Q}$ such that:

$$0 < \left|x - \frac{p}{q}\right| < \frac{1}{q^n}.$$

It follows rather quickly from Liouville's theorem that all such numbers are transcendental. Indeed, if a Liouville number x were algebraic, then by Liouville's theorem, it would be approximable to order at most $\deg(x)$. However, the above shows that x is approximable to arbitrary order. The following is a classical construction:

Theorem 3.6 ([5, Liouville Numbers]). Let $b \in \mathbb{Z}_{\geq 2}$ and consider any sequence of integers $\{a_k\}_{k=1}^{\infty}$ such that for all $k \in \mathbb{N}$, $a_k \in \{0, 1, \dots, b-1\}$ and $a_k \neq 0$ for infinitely many k . Define

$$\xi := \sum_{k=1}^{\infty} \frac{a_k}{b^{k!}}.$$

Then ξ is a Liouville number.

Proof. Note that ξ is well defined. For $k \geq 1$ we have $\frac{a_k}{b^{k!}} \leq \frac{b-1}{b^{k!}} \leq (b-1) \left(\frac{1}{b}\right)^k$. Since $\sum_{k=1}^{\infty} (b-1) \left(\frac{1}{b}\right)^k$ converges by the geometric series test, it follows by comparison that ξ does also.

Let $n \in \mathbb{N}$ and consider $q_n := b^{n!}$ and $p_n := \sum_{k=1}^n a_k b^{n!-k!}$. Note that $p_n \in \mathbb{Z}$, since $a_k b^{n!-k!} \in \mathbb{Z}$ for all $1 \leq k \leq n$. Thus, we have:

$$\begin{aligned} 0 < \left|\xi - \frac{p_n}{q_n}\right| &= \left|\xi - \sum_{k=1}^n \frac{a_k}{b^{k!}}\right| = \sum_{k=n+1}^{\infty} \frac{a_k}{b^{k!}} \leq \sum_{k=0}^{\infty} \frac{b-1}{b^{(k+(n+1))!}}, \\ &< \frac{b-1}{b^{(n+1)!}} \sum_{k=0}^{\infty} \frac{1}{b^k} = \frac{b-1}{b^{(n+1)!}} \frac{1}{1 - \frac{1}{b}} = \frac{b}{b^{(n+1)!}}, \\ &= \frac{1}{b^{n \cdot n!}} \frac{b}{b^{n!}} \leq \frac{1}{b^{n \cdot n!}} = \frac{1}{q_n^n}. \end{aligned}$$

It follows that ξ is a Liouville number. $\square$

Chapter 4

Application: Binary Homogeneous Polynomials

4.1 Overview and the Role of Liouville's Theorem

In the last chapter, we proved Liouville's Theorem, which states that no algebraic irrational number can be approximated to any order exceeding its degree. We then used this to construct explicit examples of transcendental numbers. In this chapter, we will employ similar techniques to approach a different problem. Namely, given a homogeneous polynomial $F(x, y) \in \mathbb{Z}[x, y]$, we attempt to determine whether

$$F(x, y) = k$$

admits finitely or infinitely many integer solutions, for a given non-zero integer k .

The link between homogeneous polynomials and approximability stems from the fact that any binary form¹ can be expressed as the evaluation of a univariate polynomial at x/y . More specifically, if there are infinitely many integer solutions (x, y) to an equation defined by a binary form, we obtain infinitely many rational numbers of the form x/y that approximate a root ξ of the corresponding univariate polynomial to a certain degree. In most cases, the degree condition imposed on these approximations contradicts Liouville's theorem, and we can deduce that the solution set must be finite. However, there exist two explicit families of homogeneous forms — labelled (E1) and (E2) in the following section — for which this argument cannot force a contradiction. After analysing these exceptional cases, we are left with one situation in which Liouville's theorem does not suffice. This is precisely what motivates the introduction of Thue's theorem in the next chapter.

To better understand the idea described above, we begin by examining two instructive examples:

Example 4.1. Consider the following homogeneous polynomial equation of degree 3:

$$F(x, y) := x^2y - 2y^3.$$

We restrict attention to integer pairs (x, y) with $x > 0$, $y > 0$, and $\gcd(x, y) = 1$.

¹A *binary form* is a homogeneous polynomial in two variables.

Let k be a non-zero integer. Factoring F gives

$$F(x, y) = y^3 \left(\frac{x^2}{y^2} - 2 \right).$$

Hence, the condition $F(x, y) = k$ is equivalent to

$$k = y^3 \left(\frac{x^2}{y^2} - 2 \right) \implies \frac{x^2}{y^2} - 2 = \frac{k}{y^3} \implies \left| \sqrt{2} - \frac{x}{y} \right| = \frac{|k|}{|y|^3 \left| \sqrt{2} + \frac{x}{y} \right|} \leq \frac{|k|}{|y|^3 |\sqrt{2}|}.$$

Since $\sqrt{2}$ is an algebraic number of degree 2, Liouville's theorem implies that there are only finitely many reduced rational numbers x/y with $x > 0$ and $y > 0$ satisfying the above inequality. Thus, there can only be finitely many such integer pairs (x, y) satisfying $F(x, y) = k$.

Example 4.2. Now, consider the following homogeneous binary polynomial:

$$F(x, y) := x^3 - 5y^3.$$

Again, we restrict attention to integer pairs (x, y) with $x > 0$, $y > 0$, and $\gcd(x, y) = 1$.

Then the condition $F(x, y) = 9$ is equivalent to

$$\left(\frac{x}{y} \right)^3 - 5 = \frac{9}{y^3} \implies \left| \sqrt[3]{5} - \frac{x}{y} \right| = \frac{9}{|y|^3 \left| \sqrt[3]{5}^2 + \sqrt[3]{5} \frac{x}{y} + \left(\frac{x}{y} \right)^2 \right|} \leq \frac{9}{|y|^3 |\sqrt[3]{5}^2|}.$$

Since $\sqrt[3]{5}$ is an irrational algebraic number of degree 3, Liouville's theorem does not rule out infinitely many rational approximations of order 3. Thus, unlike in the previous example, Liouville's theorem alone does not allow us to determine that there are only finitely many such integer pairs (x, y) satisfying $F(x, y) = 9$.

However, in the next chapter, we shall see that Thue's theorem² implies that this conclusion is true.

4.2 Definitions and Basic Properties

We now turn to homogeneous polynomials. They provide a natural bridge between factorisation in one variable and Diophantine equations in two variables.

Definition 4.3 (Homogeneous Polynomial). A polynomial $F(x_1, \dots, x_n)$ is called *homogeneous of degree d* if F is a linear combination of monomials of total degree d , i.e.

$$F(x_1, \dots, x_n) = \sum_{i_1 + \dots + i_n = d} a_{i_1, \dots, i_n} x_1^{i_1} \cdots x_n^{i_n} \quad \text{with } a_{i_1, \dots, i_n} \in \mathbb{Z}.$$

We shall focus on such polynomials in two variables, i.e. binary forms. The following lemma shows that any homogeneous binary form $F(x, y)$ can be rewritten as the evaluation of a univariate polynomial at x/y . Thus, when analysing the roots of such polynomials, we do not need to work with the full two-variable structure, which is considerably more complicated.

²Thue's theorem states that any algebraic number of degree $d \geq 2$ is not approximable to any order greater than $d/2 + 1$.

Lemma 4.4. *Let $F(x, y) \in \mathbb{Z}[x, y]$ be homogeneous of degree d . Then there exists $f(t) \in \mathbb{Z}[t]$ such that*

$$F(x, y) = y^d f\left(\frac{x}{y}\right).$$

In particular, writing $r = \deg(f) \leq d$ and factoring f over $\mathbb{C}$ gives

$$F(x, y) = cy^d \prod_{i=1}^L \left(\frac{x}{y} - \alpha_i\right)^{m_i} = cy^{d-r} \prod_{i=1}^L (x - \alpha_i y)^{m_i},$$

where $c \in \mathbb{C}$, the α_i are distinct, $m_i \geq 1$, and $\sum_{i=1}^L m_i = r$.

Proof. Since $F(x, y)$ is homogeneous of degree d , it can be written as

$$F(x, y) = \sum_{i=0}^d a_i x^i y^{d-i} = y^d \sum_{i=0}^d a_i \left(\frac{x}{y}\right)^i.$$

Setting $f(t) = \sum_{i=0}^d a_i t^i \in \mathbb{Z}[t]$ gives the first claim.

Let $r = \deg(f)$, so $r \leq d$. By the *Fundamental Theorem of Algebra*,

$$f(t) = c \prod_{i=1}^L (t - \alpha_i)^{m_i},$$

with distinct α_i and multiplicities m_i summing up to r . Substituting $t = x/y$ and multiplying by y^d gives us

$$F(x, y) = y^d f\left(\frac{x}{y}\right) = cy^d \prod_{i=1}^L \left(\frac{x}{y} - \alpha_i\right)^{m_i} = cy^{d-r} \prod_{i=1}^L (x - \alpha_i y)^{m_i}.$$

□

4.3 Minimal Polynomials and Multiplicity

For any algebraic number β , we can gain a lot of information by considering its associated minimal polynomial. In particular, the minimal polynomial generates all the annihilating polynomials of β . From this, we can form a relationship between the degree of the algebraic number and its multiplicity as a root of a polynomial that annihilates it. This relationship will be instrumental in our later analysis of homogeneous forms, as it will allow us to compare the multiplicities of roots of polynomials formed by the factors of a binary form and its associated homogeneous form.

The exact relationship between the degree of an algebraic number and its multiplicity as a root for a given annihilating polynomial is given by the following theorem:

Theorem 4.5. *Let β be an algebraic number with minimal polynomial $m(x) \in \mathbb{Q}[x]$. If $f(x) \in \mathbb{Z}[x]$ is an integer polynomial that satisfies $f(\beta) = 0$ with multiplicity n , then*

$$f(x) = m(x)^n h(x) \quad \text{for some } h(x) \in \mathbb{Q}[x],$$

*and in particular $\deg(\beta) = \deg(m) \leq \frac{1}{n} \deg(f)$.*³

³Here $\deg(\beta)$ denotes the degree of the minimal polynomial of β .

Proof. A full proof is given in Supplementary Material, Theorem S.1.2. $\square$

4.4 Analysis of Binary Forms

Our goal in the following sections is to determine whether $F(x, y) = k$ has finitely or infinitely many integer solutions for a fixed homogeneous polynomial F and integer $k \neq 0$. Before using the full strength of Liouville's theorem, we shall carry out a preliminary analysis of the implications of having infinitely many solutions to $F(x, y) = k$, which reduces the problem to a question about rational approximation to a root of f .

Remark 4.6 ([4]). The reason we omit $k = 0$ is that integer solutions to $F(x, y) = 0$ can be read off directly from the factorisation in Lemma 4.4. Indeed, we may write $F(x, y)$ as:

$$F(x, y) = cy^{d-r} \prod_{i=1}^L (x - \alpha_i y)^{m_i},$$

where we see that $F(x, y) = 0$ holds if and only if one of the factors vanishes. In particular:

- If $d > r$, then the factor y^{d-r} forces $F(x, 0) = 0$ for every $x \in \mathbb{Z}$, and hence there are infinitely many solutions of the form $(x, 0)$.
- If $d = r$ and $\alpha_i = 0$ for some i , then the factor $(x - \alpha_i y)^{m_i} = x^{m_i}$, which implies $F(0, y) = 0$ for every $y \in \mathbb{Z}$. Thus, there are infinitely many solutions of the form $(0, y)$.
- If $d = r$, each α_i is non-zero and $\alpha_i = p/q \in \mathbb{Q}$ for some i , then (pt, qt) is a solution for each $t \in \mathbb{Z}$. This gives infinitely many integer solutions (x, y) satisfying $x - \alpha_i y = 0$.
- If $d = r$ and each non-zero α_i is irrational, then $x - \alpha_i y = 0$ has no non-trivial integer solutions, for each i . Thus, $F(x, y) = 0$ has no non-trivial integer solutions.

Therefore, $F(x, y) = 0$ has infinitely many integer solutions unless $d = r$ and every non-zero root α_i is irrational.

The following lemmas give some preliminary restrictions on the solutions to $F(x, y) = k$ when $k \neq 0$. The first one shows that there are only finitely many solutions with $xy = 0$. This allows us to subsequently shift our focus to the case where $xy \neq 0$.

Lemma 4.7. *Let $F(x, y) \in \mathbb{Z}[x, y]$ be a homogeneous polynomial of degree d , and let $k \in \mathbb{Z} \setminus \{0\}$. Then $F(x, y) = k$ has only finitely many integer solutions (x, y) when $xy = 0$.*

Proof. Write

$$F(x, y) = \sum_{i=0}^d a_i x^i y^{d-i}$$

with $a_i \in \mathbb{Z}$ and suppose $xy = 0$.

If $y = 0$, then $F(x, 0) = a_d x^d$. Hence any solution to $F(x, 0) = k$ must satisfy $a_d x^d = k$. Since $k \neq 0$, there are only finitely many such integers x .

If $x = 0$, then $F(0, y) = a_0 y^d$. Hence any solution to $F(0, y) = k$ must satisfy $a_0 y^d = k$, and again there are only finitely many such integers y .

Therefore $F(x, y) = k$ has only finitely many integer solutions with $xy = 0$. $\square$

From now on, we restrict our attention to the case where $xy \neq 0$. The next lemma shows that, in this situation, the rational number x/y determines the solution up to sign. This observation will be essential when we later pass to rational approximations.

Lemma 4.8. *Let $k \in \mathbb{Z} \setminus \{0\}$ and let $F(x, y) \in \mathbb{Z}[x, y]$ be homogeneous of degree $d \geq 1$. If $(x, y) \in \mathbb{Z}^2$ satisfies $F(x, y) = k$ with $xy \neq 0$, and if (x', y') is another integer solution with $\frac{x'}{y'} = \frac{x}{y}$, then $(x', y') = (x, y)$ or $(x', y') = -(x, y)$. In other words, the rational number x/y uniquely determines the solution up to sign.*

Proof. If $\frac{x'}{y'} = \frac{x}{y}$, then there exists some non-zero rational number t such that $(x', y') = (tx, ty)$. Since F is homogeneous of degree $d > 0$,

$$F(x', y') = F(tx, ty) = t^d F(x, y) = t^d k.$$

But also $F(x', y') = k$, and $k \neq 0$, so $t^d = 1$. As $t \in \mathbb{Q}$, this forces $t = \pm 1$. Hence

$$(x', y') = (x, y) \quad \text{or} \quad (x', y') = -(x, y).$$

□

4.5 Constructing Rational Approximations

Using the tools developed so far, we can translate the question of whether $F(x, y) = k$ has infinitely many solutions into a question about how well a root of the associated polynomial $f(x/y)$ can be approximated by rational numbers.

Proposition 4.9. *Let $F(x, y) \in \mathbb{Z}[x, y]$ be a homogeneous polynomial of degree d , and fix $k \in \mathbb{Z} \setminus \{0\}$. By Lemma 4.4 we can write*

$$F(x, y) = cy^d \prod_{i=1}^L \left(\frac{x}{y} - \alpha_i \right)^{m_i},$$

where $c \in \mathbb{C}$, the α_i are distinct, $m_i \geq 1$, and $\sum_{i=1}^L m_i = r \leq d$.

Suppose the equation $F(x, y) = k$ has infinitely many integer solutions with $xy \neq 0$. Then there exist an index i and a constant $K(\alpha_i) > 0$ such that infinitely many integer solutions (x, y) satisfy

$$\left| \alpha_i - \frac{x}{y} \right| \leq \frac{K(\alpha_i)}{|y|^{d/m_i}}.$$

Proof. Let $S = \{(x, y) \in \mathbb{Z}^2 : F(x, y) = k, xy \neq 0\}$, which is infinite by assumption. By Lemma 4.4, $F(x, y) = cy^d \prod_{i=1}^L (x/y - \alpha_i)^{m_i}$. If $L = 1$, then f has only one distinct root, say $\alpha = \alpha_1$, and the conclusion follows immediately from

$$|k| = |c| |y|^d \left| \frac{x}{y} - \alpha \right|^{m_1}.$$

Hence we may assume that $L \geq 2$. Let us partition S according to the nearest root:

$$S_i = \left\{ (x, y) \in S : \left| \frac{x}{y} - \alpha_i \right| \leq \left| \frac{x}{y} - \alpha_j \right| \text{ for all } j \neq i \right\}.$$

Then $S = \bigcup_{i=1}^L S_i$, so at least one of these S_i is infinite. Without loss of generality, suppose S_1 is infinite and let $\alpha = \alpha_1$.

Let

$$\varepsilon = \frac{1}{2} \min_{2 \leq j \leq L} |\alpha - \alpha_j| > 0.$$

We split S_1 into two subsets as follows:

$$S'_1 = \left\{ (x, y) \in S_1 : \left| \frac{x}{y} - \alpha \right| < \varepsilon \right\}, \quad S''_1 = S_1 \setminus S'_1.$$

If $(x, y) \in S''_1$, then $|x/y - \alpha| \geq \varepsilon$, and by the definition of S_1 ,

$$\left| \frac{x}{y} - \alpha_j \right| \geq \left| \frac{x}{y} - \alpha \right| \geq \varepsilon \quad \text{for all } j \neq 1.$$

Hence

$$|k| = |c||y|^d \prod_{i=1}^L \left| \frac{x}{y} - \alpha_i \right|^{m_i} \geq |c||y|^d \varepsilon^r,$$

where $r = \sum m_i$. Thus $|y|^d \leq \frac{|k|}{|c|\varepsilon^r}$, so y is bounded. For each fixed y , the equation $F(x, y) = k$ has only finitely many solutions in x , hence S''_1 is finite.

Since S_1 is infinite, it follows that S'_1 is infinite. For $(x, y) \in S'_1$ we have

$$\left| \frac{x}{y} - \alpha_j \right| \geq |\alpha - \alpha_j| - \left| \frac{x}{y} - \alpha \right| > 2\varepsilon - \varepsilon = \varepsilon \quad \text{for all } j \neq 1.$$

Therefore

$$|k| = |c||y|^d \left| \frac{x}{y} - \alpha \right|^{m_1} \prod_{j=2}^L \left| \frac{x}{y} - \alpha_j \right|^{m_j} \geq |c||y|^d \left| \frac{x}{y} - \alpha \right|^{m_1} \varepsilon^{r-m_1}.$$

Rearranging,

$$\left| \alpha - \frac{x}{y} \right|^{m_1} \leq \frac{|k|}{|c|\varepsilon^{r-m_1}} \frac{1}{|y|^d}.$$

Hence

$$\left| \alpha - \frac{x}{y} \right| \leq \frac{K(\alpha)}{|y|^{d/m_1}}, \quad \text{where } K(\alpha) = \left(\frac{|k|}{|c|\varepsilon^{r-m_1}} \right)^{1/m_1} > 0.$$

Since S'_1 is infinite, the inequality holds for infinitely many (x, y) . □

Moreover, the root α in Proposition 4.9 must be real, as shown by the following lemma:

Lemma 4.10. *In the setting of Proposition 4.9, the corresponding root α must be real.*

Proof. By Proposition 4.9, there exist infinitely many integer solutions (x, y) such that

$$\left| \alpha - \frac{x}{y} \right| \leq \frac{K(\alpha)}{|y|^{d/m_1}} \quad \text{for some constant } K(\alpha) > 0. \quad (4.1)$$

Suppose that α is not real. Let $\alpha = \lambda + \mu i$, where $\lambda, \mu \in \mathbb{R}$ and $\mu \neq 0$. Then for any real number x/y we have

$$\left| \alpha - \frac{x}{y} \right|^2 = \left| \lambda - \frac{x}{y} + \mu i \right|^2 = \left(\lambda - \frac{x}{y} \right)^2 + \mu^2 \geq \mu^2.$$

Hence $|\alpha - x/y| \geq |\mu|$. Combining this with (4.1) gives

$$|\mu| \leq \left| \alpha - \frac{x}{y} \right| \leq \frac{K(\alpha)}{|y|^{d/m_1}} \implies |y|^{d/m_1} \leq \frac{K(\alpha)}{|\mu|}.$$

Therefore y is bounded. For each fixed y , the equation $F(x, y) = k$ admits only finitely many integer solutions in x , so there are only finitely many such pairs (x, y) altogether. This contradicts the infinitude obtained in Proposition 4.9. Hence α must be real. $\square$

4.6 Two Exceptional Families

We can now use Proposition 4.9 to classify homogeneous polynomials in two variables. For all but two forms that F can take, we shall see that assuming the existence of infinitely many integer solutions to $F(x, y) = k$ leads to a contradiction, by virtue of Liouville's theorem.

We first establish a simple lemma concerning the form of the minimal polynomial of a real algebraic number of degree 2. We will use this to identify one of the two exceptional families of binary forms whose analysis is not amenable to Liouville's theorem.

Lemma 4.11. *Let $\alpha \in \mathbb{R}$ be an algebraic number of degree 2, and write its minimal polynomial as*

$$g(t) = at^2 + bt + c \quad (a, b, c \in \mathbb{Z}, a \neq 0).$$

Then the discriminant $\Delta = b^2 - 4ac$ is strictly positive and not a perfect square.

Proof. Using the quadratic formula, the roots of g are

$$\frac{-b \pm \sqrt{\Delta}}{2a}, \quad \Delta = b^2 - 4ac.$$

As α is real, we must have $\Delta \geq 0$. If $\Delta = 0$, then $\alpha = -\frac{b}{2a} \in \mathbb{Q}$, contradicting $\deg(\alpha) = 2$. Hence $\Delta > 0$.

If Δ were a perfect square, then $\sqrt{\Delta} \in \mathbb{Z}$ and both roots would be rational, again contradicting $\deg(\alpha) = 2$. Thus $\Delta > 0$ and Δ is not a perfect square. $\square$

We can now state the main reduction result of this chapter. For convenience we single out two special forms:

(E1) $F(x, y) = \lambda(ax + by)^M$ for some $\lambda \neq 0$, a, b not both zero, and $M \in \mathbb{N}$.

(E2) $F(x, y) = \lambda(ax^2 + bxy + cy^2)^M$ for some $\lambda \neq 0$, $\Delta = b^2 - 4ac > 0$ not a perfect square, and $M \in \mathbb{N}$.

The following theorem shows that, apart from these two explicit families, any equation $F(x, y) = k$ with infinitely many integer solutions forces rational approximations that are incompatible with Liouville's theorem, unless there exists a real root α of f with $\deg(\alpha) \geq 3$. For such roots, Liouville's theorem is not strong enough to provide a contradiction.

Theorem 4.12. *Let $F(x, y) \in \mathbb{Z}[x, y]$ be a homogeneous polynomial of degree $d \geq 1$, and let $k \in \mathbb{Z} \setminus \{0\}$. Suppose that the equation*

$$F(x, y) = k$$

has infinitely many integer solutions. Then one of the following holds:

1. F belongs to one of the families (E1) or (E2) above;
2. After factorising $F = y^d f(x/y)$ as in Lemma 4.4, there exists a real root α of f with $\deg(\alpha) \geq 3$ and multiplicity m_0 , and a constant $K(\alpha) > 0$, such that infinitely many integer solutions (x, y) of $F(x, y) = k$ satisfy

$$\left| \alpha - \frac{x}{y} \right| \leq \frac{K(\alpha)}{|y|^{d/m_0}}.$$

In particular, if the second case can be ruled out (as will be done in the next chapter using Thue's theorem), then the only cases in which $F(x, y) = k$ admits infinitely many integer solutions are (E1) and (E2).

Proof. By Lemma 4.7, there are only finitely many solutions with $xy = 0$. Hence, we may restrict our attention to $xy \neq 0$. Let us factor $F(x, y) = y^d f(x/y)$ as in Lemma 4.4, where

$$f(t) = c \prod_{i=1}^L (t - \alpha_i)^{m_i}, \quad \sum_{i=1}^L m_i = r = \deg(f) \leq d,$$

where the α_i are distinct complex numbers. By Proposition 4.9 and Lemma 4.10, there exists a real root $\alpha = \alpha_1 \in \mathbb{R}$ with multiplicity $m_0 := m_1$ and a constant $K(\alpha) > 0$ such that infinitely many solutions (x, y) of $F(x, y) = k$ satisfy

$$\left| \alpha - \frac{x}{y} \right| \leq \frac{K(\alpha)}{|y|^{d/m_0}}. \quad (4.2)$$

Let $g(t) \in \mathbb{Q}[t]$ be the minimal polynomial of α and set $n = \deg(\alpha) = \deg(g)$. Since α is a root of f of multiplicity at least m_0 , Theorem 4.5 gives

$$f(t) = g(t)^{m_0} h(t) \quad \text{for some } h(t) \in \mathbb{Q}[t].$$

Comparing degrees yields $r = \deg(f) \geq m_0 \deg(g) = m_0 n$, and thus $\frac{r}{m_0} \geq n$. Since $r \leq d$, we also have

$$\frac{d}{m_0} \geq \frac{r}{m_0} \geq n. \quad (4.3)$$

We now distinguish cases according to $n = \deg(\alpha)$.

Case I: $n \geq 3$. From (4.3) we have $\frac{d}{m_0} \geq n \geq 3$. Hence (4.2) produces infinitely many rational approximations to α with exponent at least n , where α is algebraic of degree $n \geq 3$:

$$\left| \alpha - \frac{x}{y} \right| \leq \frac{K(\alpha)}{|y|^{d/m_0}} \leq \frac{K(\alpha)}{|y|^n}.$$

This is exactly the alternative described in the second case of the theorem, and it does not contradict Liouville's theorem, since the latter only gives an upper bound of n for the exponent of rational approximations to α .

Case II: $n = 2$. Then α is a real quadratic irrational. We consider two subcases.

IIa: $d > r$. Since $r \geq 2m_0$, we have $\frac{d}{m_0} > \frac{r}{m_0} \geq 2$. Thus (4.2) gives infinitely many rational approximations of order strictly greater than 2 to the quadratic irrational α , contradicting Liouville's theorem. Hence this case cannot occur.

IIb: $d = r$. Then $\frac{d}{m_0} = \frac{r}{m_0} \geq 2$. If $\frac{r}{m_0} > 2$, we obtain the same contradiction as in **IIa**, so the only remaining possibility is $\frac{r}{m_0} = 2$. Equivalently, $r = 2m_0$. Taking degrees

in $f(t) = g(t)^{m_0}h(t)$ gives

$$2m_0 = \deg(f) = 2m_0 + \deg(h),$$

hence $\deg(h) = 0$ and therefore $f(t) = \lambda g(t)^{m_0}$ for some non-zero constant λ . Writing

$$g(t) = at^2 + bt + c \quad (a, b, c \in \mathbb{Z}, a \neq 0),$$

and homogenising, we obtain

$$F(x, y) = y^{2m_0} f(x/y) = \lambda(ax^2 + bxy + cy^2)^{m_0}.$$

Since α is quadratic irrational, the discriminant $\Delta = b^2 - 4ac > 0$ is not a perfect square. Hence F is of the explicit type (E2).

Case III: $n = 1$. Then $\alpha \in \mathbb{Q}$. Write $\alpha = p/q$ in lowest terms. Again we split into two subcases.

IIIa: $d > r$. Since $r \geq m_0$, we have $\frac{d}{m_0} > \frac{r}{m_0} \geq 1$. Thus (4.2) yields infinitely many rational approximations of order strictly greater than 1 to a rational number. This forces $x/y = \alpha$ for all sufficiently large $|y|$, since if $\alpha = p/q$ in lowest terms and $x/y \neq \alpha$, then $\left| \alpha - \frac{x}{y} \right| \geq \frac{1}{q|y|}$. By Lemma 4.8 there can be only finitely many such integer solutions, contradicting infinitude. Hence this case cannot occur.

IIIb: $d = r$. Then $\frac{d}{m_0} = \frac{r}{m_0} \geq 1$. If $\frac{r}{m_0} > 1$, we obtain the same contradiction as above, so the only possibility is $r = m_0$. In this case f has exactly one root of multiplicity m_0 , hence $f(t) = \lambda(t - \alpha)^{m_0}$ for some non-zero constant λ . Writing $\alpha = -b/a$ and homogenising gives

$$F(x, y) = \lambda(ax + by)^{m_0},$$

which is precisely the explicit type (E1).

We conclude that if infinitely many integer solutions exist, then either F is of one of the explicit types (E1) or (E2), or else Case I occurs, which is exactly the alternative described in the second case of the theorem. $\square$

Theorem 4.12 provides the only impediment to a purely Liouville proof of finiteness of solutions to this problem. In fact, if the exceptional cases (E1) and (E2) are set aside, infinitely many solutions to $F(x, y) = k$ would force the existence of a real algebraic number α of degree $n \geq 3$ that admits infinitely many rational approximations x/y with an exponent at least $d/m_0 \geq n$; cf. (4.2) and (4.3).

Liouville's theorem cannot yield a contradiction here, since for an algebraic number of degree n it only rules out infinitely many rational approximations of exponent strictly greater than n .

The next chapter introduces Thue's theorem, which strengthens Liouville's result. Thue's theorem states that every algebraic number of degree $n \geq 2$ is not approximable to any order greater than $n/2 + 1$. Applying Thue's theorem to inequalities of the form (4.2) eliminates the remaining cases with $n \geq 3$. Once Thue's result is established in the next chapter, we can complete our argument, concluding that the only homogeneous polynomials F for which the equation $F(x, y) = k$ can have infinitely many integer solutions, with $k \neq 0$, are those of the forms (E1) and (E2).

Chapter 5

Thue's Theorem

5.1 Preliminaries for Thue's Theorem

In this chapter, we prove Thue's theorem, which states that if β is an algebraic number of degree $d \geq 2$, then β is not approximable to any order greater than $\frac{d}{2} + 1$.

To prove this result, Thue considered polynomials in two variables of the form $P(x, y) = P_1(x)y + P_0(x) \in \mathbb{Z}[x, y]$. Recall that in the proof of Liouville's theorem, we utilised explicit rational approximations to obtain a non-zero lower bound on a given univariate polynomial. However, when considering polynomials of the aforementioned form, the existence of such approximations cannot be guaranteed. To overcome this, we instead take the j -th partial derivative of P with respect to x , where j is bounded and $\partial_x^j P(r_1, r_2) \neq 0$. This yields a new polynomial $\tilde{P}$ satisfying $\tilde{P}(r_1, r_2) \neq 0$. Moreover, by construction, $\tilde{P}$ inherits many of the properties of P .

The following lemma will play a key role in the construction of a polynomial with the desired properties:

Lemma 5.1 (Siegel's Lemma [6, Lemma 6.1]). *Let M, N be positive integers with $M < N$. Consider a system*

$$\sum_{j=1}^N a_{ij}x_j = 0 \quad \text{for } i = 1, \dots, M$$

of M homogeneous linear equations in N unknowns. If there exists $B \geq 1$ such that each a_{ij} is an integer satisfying $|a_{ij}| < B$, then there exists a non-zero integer solution $(x_1, x_2, \dots, x_N) \in \mathbb{Z}^N \setminus \{\mathbf{0}\}$ of the system such that

$$\max_{1 \leq i \leq N} |x_i| \leq 2(3NB)^{\frac{M}{N-M}}.$$

Proof. Let A be the $M \times N$ matrix with entries a_{ij} . Let $S = \{(x_1, x_2, \dots, x_N) \in \mathbb{Z}^N : 0 \leq x_i \leq X \text{ for } 1 \leq i \leq N\}$ be the set of integer lattice points in $[0, X]^N$, where X is a positive integer to be chosen later. Then the number of elements in S is $(X+1)^N$.

For each $\mathbf{x} \in S$, we have

$$|(A\mathbf{x})_j| = \left| \sum_{i=1}^N a_{ji}x_i \right| \leq \sum_{i=1}^N |a_{ji}||x_i| \leq NBX.$$

for $j = 1, 2, \dots, M$. Hence, $(A\mathbf{x})_j \in [-NBX, NBX]$, and therefore the vector $A\mathbf{x}$ takes at most $T = (2NBX + 1)^M$ different values.

Now consider the map $f : S \rightarrow \mathbb{Z}^M$ defined by $f(\mathbf{x}) = A\mathbf{x}$. If $|S| > T$, then by the *pigeonhole principle*, there exist distinct $\mathbf{x}, \mathbf{y} \in S$ such that $f(\mathbf{x}) = f(\mathbf{y})$. Hence $A(\mathbf{x} - \mathbf{y}) = 0$, and $\mathbf{x} - \mathbf{y}$ is a non-trivial integer solution of the system. Since $\mathbf{x}, \mathbf{y} \in S$, we have $|x_i - y_i| \leq X$ for all i , and therefore $\max_{1 \leq i \leq N} |x_i - y_i| \leq X$.

Thus, it suffices to choose X such that

$$(X + 1)^N > (2NBX + 1)^M.$$

Since $M < N$, take $X = \lceil (3NB)^{\frac{M}{N-M}} \rceil \geq (3NB)^{\frac{M}{N-M}}$. Then $X^{N-M} \geq (3NB)^M$, and hence $X^N \geq (3NBX)^M$. Therefore, for this choice of X , we have

$$(X + 1)^N > X^N \geq (3NBX)^M \geq (2NBX + 1)^M,$$

where the last inequality holds because $NBX \geq 1$. It follows that there exists a non-trivial integer solution $\mathbf{z} = \mathbf{x} - \mathbf{y}$ such that

$$\max_{1 \leq i \leq N} |z_i| = \max_{1 \leq i \leq N} |x_i - y_i| \leq X \leq 2(3NB)^{\frac{M}{N-M}}.$$

□

In the proof of Liouville's theorem, the multiplicity condition imposed on the algebraic number resulted in a system of linear equations. To guarantee non-zero integer solutions to this system, we set the degree of the polynomial equal to the the product of the degree of β and its multiplicity as a root. Siegel's lemma will allow us to find a non-trivial, integer-valued solution to a similar system that is bounded from above. This leaves us with a non-zero polynomial whose integer coefficients satisfy an explicit upper bound. After constructing our polynomial, we establish a lower bound on its height.

Proposition 5.2 ([2, Proposition 2.1]). *Let β be an algebraic number. For any integer $m \geq 1$ and any $\varepsilon > 0$ sufficiently small, there exists a non-zero polynomial*

$$P(x, y) = P_1(x)y + P_0(x) \in \mathbb{Z}[x, y]$$

such that the following three properties hold:

- (a) $\partial_x^j P(\beta, \beta) = 0$ for all $j = 0, 1, \dots, m - 1$
- (b) $|P| \leq B(\beta, \varepsilon)^m$
- (c) $\deg(P_0), \deg(P_1) \leq (1 + \varepsilon) \frac{\deg(\beta)}{2} m$

Proof. Consider a polynomial $P(x, y) = P_1(x)y + P_0(x)$, where $P_0(x) = \sum_{i=0}^D a_i x^i$ and $P_1(x) = \sum_{i=0}^D b_i x^i$, with $a_i, b_i \in \mathbb{Z}$. We shall choose the degree D later. We require that

$$\partial_x^j P(\beta, \beta) = 0 \quad \text{for } j = 0, 1, \dots, m - 1.$$

This gives the following m linear equations, where the a_i and b_i are treated as unknowns:

$$\frac{1}{j!} \partial_x^j P(\beta, \beta) = \sum_{i=j}^D b_i \binom{i}{j} \beta^{i-j+1} + \sum_{i=j}^D a_i \binom{i}{j} \beta^{i-j} = 0 \quad \text{for } j = 0, 1, \dots, m - 1.$$

Let $m_\beta(x) = \sum_{i=0}^{d_0} q_i x^i \in \mathbb{Z}[x]$ denote the minimal polynomial of β , where $d_0 = \deg(\beta)$. By Lemma 3.2, every power β^d can be expressed as a linear combination of $1, \beta, \dots, \beta^{d_0-1}$.

That is,

$$\beta^d = \frac{1}{q_{d_0}^d} \sum_{k=0}^{d_0-1} A_{k,d} \beta^k \quad \text{with} \quad A_{k,d} \in \mathbb{Z} \quad \text{and} \quad |A_{k,d}| \leq [2|m_\beta|]^d.$$

Substituting these expressions into the above equations, we obtain

$$0 = \sum_{i=j}^D b_i \binom{i}{j} \left(\sum_{k=0}^{d_0-1} \frac{A_{k,i-j+1}}{q_{d_0}^{i-j+1}} \beta^k \right) + \sum_{i=j}^D a_i \binom{i}{j} \left(\sum_{k=0}^{d_0-1} \frac{A_{k,i-j}}{q_{d_0}^{i-j}} \beta^k \right).$$

Multiplying through by $q_{d_0}^{D+1}$ and rearranging the sums gives

$$\sum_{k=0}^{d_0-1} \left(\sum_{i=j}^D b_i \binom{i}{j} A_{k,i-j+1} q_{d_0}^{D-(i-j)} + \sum_{i=j}^D a_i \binom{i}{j} A_{k,i-j} q_{d_0}^{D+1-(i-j)} \right) \beta^k = 0.$$

This is a linear combination of $1, \beta, \beta^2, \dots, \beta^{d_0-1}$ with integer coefficients. Indeed, $A_{k,d} \in \mathbb{Z}$, and the exponents $D - (i - j)$ and $D + 1 - (i - j)$ are non-negative since $i \leq D$. Since $\{1, \beta, \beta^2, \dots, \beta^{d_0-1}\}$ is linearly independent over $\mathbb{Q}$, it follows that

$$\sum_{i=j}^D b_i \tilde{A}_{i,j,k} + \sum_{i=j}^D a_i \tilde{B}_{i,j,k} = 0$$

for all $j = 0, 1, \dots, m-1$ and $k = 0, 1, \dots, d_0-1$, where $\tilde{A}_{i,j,k} = \binom{i}{j} A_{k,i-j+1} q_{d_0}^{D-(i-j)}$ and $\tilde{B}_{i,j,k} = \binom{i}{j} A_{k,i-j} q_{d_0}^{D+1-(i-j)}$ are both integers.

Thus, we obtain a homogeneous system of md_0 linear equations in $2D + 2$ unknowns over $\mathbb{Z}$. By Lemma 5.1, it suffices to have $2D + 2 > md_0$, or equivalently, $D > \frac{d_0}{2}m - 1$. If we choose $D = \lfloor (1 + \varepsilon) \frac{d_0}{2}m \rfloor$ for any $\varepsilon > 0$, then there exists a non-zero polynomial $P(x, y) \in \mathbb{Z}[x, y]$ satisfying property (a).

We now verify property (b). By the bound on $A_{k,d}$, we have

$$|\tilde{A}_{i,j,k}|, |\tilde{B}_{i,j,k}| \leq \binom{D}{j} [2|m_\beta|]^{i-j+1} |q_{d_0}|^{D+1-(i-j)} \leq 2^D [2|m_\beta|]^{D+1} |q_{d_0}|^{D+1} =: C(\beta)^D$$

where we used the fact that $\binom{D}{j} \leq 2^D$, $D \geq i - j$ and $|A_{k,d}| \leq [2|m_\beta|]^d$. By Lemma 5.1, the non-trivial integer solution satisfies

$$\max_{0 \leq i \leq D} \{|a_i|, |b_i|\} \leq 2(3(2D + 2)C(\beta)^D)^{\frac{md_0}{2D+2-md_0}}.$$

With our choice of D , we have $D \geq (1 + \varepsilon) \frac{d_0}{2}m - 1$, and therefore

$$\frac{md_0}{2D + 2 - md_0} \leq \frac{md_0}{2((1 + \varepsilon) \frac{d_0}{2}m - 1) + 2 - md_0} = \frac{md_0}{\varepsilon md_0} = \frac{1}{\varepsilon}.$$

It follows that

$$\max_{0 \leq i \leq D} \{|a_i|, |b_i|\} \leq 2(3(2D + 2)C(\beta)^D)^{1/\varepsilon} = 2 \cdot 3^{1/\varepsilon} (2D + 2)^{1/\varepsilon} C(\beta)^{D/\varepsilon}.$$

Since $D \leq (1 + \varepsilon) \frac{d_0}{2}m$, we have $C(\beta)^{D/\varepsilon} \leq C(\beta)^{(1+\varepsilon) \frac{d_0}{2\varepsilon}m}$. Also, $2D + 2 \leq (1 + \varepsilon)d_0m + 2$. For any fixed $k > 0$, there exists a constant $C_k > 0$ such that $x^k \leq C_k 2^x$ for all $x >$

0¹. Applying this with $k = 1/\varepsilon$ and $x = 2D + 2$, we get $(2D + 2)^{1/\varepsilon} \leq C_{1/\varepsilon} 2^{2D+2} \leq C_{1/\varepsilon} 2^{(1+\varepsilon)d_0m+2}$. Therefore,

$$\max_{0 \leq i \leq D} \{|a_i|, |b_i|\} \leq 2 \cdot 3^{1/\varepsilon} C_{1/\varepsilon} 2^{(1+\varepsilon)d_0m+2} C(\beta)^{(1+\varepsilon)\frac{d_0}{2\varepsilon}m} \leq B(\beta, \varepsilon)^m$$

for some constant $B(\beta, \varepsilon) > 0$ depending only on β and ε . This proves property (b). Finally, property (c) holds by the choice of D . This completes the proof of Proposition 5.2. $\square$

5.2 Bounding the Height of the Polynomial

In the previous section we obtained a non-zero integer polynomial, with an upper bound on its height and two properties analogous to those satisfied by the single-variable polynomials used in the proof of Liouville's theorem. We shall now establish a lower bound for the height of such a polynomial at a given rational point. To do so, we introduce the notion of the *content* of a polynomial and then prove a useful corollary of Gauss's Lemma. These results will allow us to derive a lower bound.

5.2.1 Preliminaries: Content and Gauss's Lemma

Definition 5.3 ([7, Content of a Polynomial]). Let $f(x) := a_d x^d + a_{d-1} x^{d-1} + \dots + a_0 \in \mathbb{Z}[x]$. The *content* $c(f)$ of f is defined by

$$c(f) := \gcd(a_d, a_{d-1}, \dots, a_0).$$

We say that f is *primitive* if $c(f) = 1$.

Theorem 5.4 (Gauss's Lemma [7]). *The product of two primitive polynomials is primitive.*

Proof. The proof may be found in [7]. For completeness, we include it in Supplementary Material S.1.4. $\square$

Proposition 5.5 ([7, Content of Product of Polynomials]). *Let $f(x), g(x) \in \mathbb{Z}[x]$ be two polynomials. Then we have:*

$$c(fg) = c(f) \cdot c(g)$$

Proof. The proof is presented in Supplementary Material S.1.5. $\square$

With these preliminaries in place, we can prove a useful corollary, which will help us derive a lower bound for the height of the polynomial.

Lemma 5.6 ([4, Corollary of Gauss's Lemma]). *Let $r \in \mathbb{Q} \setminus \{0\}$ and let $Q(x)$ be a non-zero polynomial in $\mathbb{Z}[x]$ such that*

$$Q(x) = \sum_i a_i x^i \in \mathbb{Z}[x] \quad \text{and} \quad Q^{(j)}(r) = 0 \text{ for } j = 0, 1, \dots, l-1, \quad l \geq 1$$

Then we have: $|Q| \geq \|r\|^l$, where $\|r\| = \max(|p|, |q|)$ for $r = \frac{p}{q}$ in reduced form.

¹Consider $f(x) = x^k 2^{-x} = x^k e^{-x \log 2}$ for $x > 0$. Then $f'(x) = x^{k-1} e^{-x \log 2} (k - x \log 2)$, so f attains its maximum at $x = \frac{k}{\log 2}$. Hence $x^k 2^{-x} \leq \left(\frac{k}{\log 2}\right)^k 2^{-k/\log 2} = \left(\frac{k}{e \log 2}\right)^k$, and thus $x^k \leq \left(\frac{k}{e \log 2}\right)^k 2^x$.

Proof. Since $Q(x) \in \mathbb{Z}[x]$ is non-zero, its content satisfies $c(Q) \geq 1$. Let $Q(x) = c(Q) \cdot \tilde{Q}(x)$ where $\tilde{Q}(x)$ is primitive. Then $|Q| = c(Q) \cdot |\tilde{Q}| \geq |\tilde{Q}|$. Therefore, it suffices to prove the lemma for the primitive polynomial $\tilde{Q}(x)$.

We shall now take the Taylor expansion of $\tilde{Q}(x)$ around $r := \frac{p}{q}$, where p and q are coprime. Since $\tilde{Q}^{(j)}(r) = 0$ for $j = 0, 1, \dots, l-1$, we obtain:

$$\tilde{Q}(x) = \left(x - \frac{p}{q}\right)^l \cdot \sum_{j=l}^D \frac{\tilde{Q}^{(j)}(r)}{j!} \left(x - \frac{p}{q}\right)^{j-l} = (qx - p)^l \cdot \frac{1}{q^l} \cdot R(x)$$

where $R(x) = \sum_{j=l}^D \frac{\tilde{Q}^{(j)}(r)}{j!} (x - \frac{p}{q})^{j-l}$. If we multiply $R(x)$ by $q^{D-l}D!$ to clear the denominators, we obtain

$$P(x) := (q^{D-l}D!)R(x) = \sum_{j=l}^D \frac{D!}{j!} \tilde{Q}^{(j)}(r) q^{D-j} (qx - p)^{j-l}.$$

Then necessarily $P(x) \in \mathbb{Z}[x]$. If we let $a = (q^{D-l}D!)$ and divide $P(x)$ by its content, say $c > 0$, then $\tilde{P}(x) = \frac{P(x)}{c} \in \mathbb{Z}[x]$ is primitive. Thus, we have:

$$\tilde{Q}(x) = (qx - p)^l \cdot \frac{1}{q^l} \cdot \frac{c}{a} \cdot \tilde{P}(x)$$

which gives

$$aq^l \cdot \tilde{Q}(x) = c \cdot (qx - p)^l \cdot \tilde{P}(x)$$

By Gauss's Lemma, if we take the content of both sides, and use the multiplicative property of the content (Proposition 5.5), we have:

$$c(aq^l \cdot \tilde{Q}) = c(c \cdot (qx - p)^l \cdot \tilde{P}) \implies |a| \cdot |q|^l \cdot c(\tilde{Q}) = |c| \cdot c((qx - p)^l) \cdot c(\tilde{P}).$$

Since $\tilde{Q}$ and $\tilde{P}$ are primitive, we have $c(\tilde{Q}) = c(\tilde{P}) = 1$. Also, since p and q are coprime, $c((qx - p)^l) = 1$. Hence, we obtain $|a| \cdot |q|^l = |c|$ and therefore:

$$\tilde{Q}(x) = \pm (qx - p)^l \cdot \tilde{P}(x)$$

Finally, we have:

$$|\tilde{Q}| \geq |(qx - p)^l| \geq \max(|p|^l, |q|^l) = \|r\|^l.$$

□

5.2.2 Lower Bounds for Height

Using the above definition of the content of a polynomial, together with a crucial corollary of Gauss's Lemma — Lemma 5.6 — we can control the height of a non-zero polynomial with a prescribed multiplicity at a given rational point. From this, we can derive a lower bound for the height of the polynomial $P(x, y)$ constructed in Proposition 5.2.

Proposition 5.7 ([4]). *Given any $\mathbf{r} = (r_1, r_2) \in \mathbb{Q}^2 \setminus \{\mathbf{0}\}$, let $P(x, y) = P_1(x)y + P_0(x) \in \mathbb{Z}[x, y]$ be a primitive non-zero integer polynomial such that*

$$\partial_x^j P(\mathbf{r}) = 0 \quad \text{for } j = 0, 1, \dots, l-1 \quad \text{where } l \geq 1.$$

Then, we have:

$$|P| \geq \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right)$$

where $D = \max(\deg(P_0), \deg(P_1))$ and $\|r\| = \max(|p|, |q|)$ for $r = \frac{p}{q}$ in reduced form.

The proof of Proposition 5.7 is not entirely straightforward. Nevertheless, the assumption $\partial_x^j P(\mathbf{r}) = 0$ for $j = 0, 1, \dots, l-1$, together with the particularly simple structure of the bivariate polynomial P , naturally prompts us to investigate what information can be deduced from this condition alone.

Expanding the derivative explicitly, we observe that for each $j = 0, 1, \dots, l-1$,

$$\partial_x^j P(\mathbf{r}) = P_1^{(j)}(r_1)r_2 + P_0^{(j)}(r_1) = 0.$$

If we define $x_j := P_1^{(j)}(r_1)$ and $y_j := P_0^{(j)}(r_1)$, then each pair (x_j, y_j) satisfies the linear condition $x_j r_2 + y_j = 0$. In other words, all vectors (x_j, y_j) lie on the same line in $\mathbb{R}^2$.

Consequently, for any two indices j_1, j_2 , the vectors (x_{j_1}, y_{j_1}) and (x_{j_2}, y_{j_2}) are linearly dependent. This implies that the determinant of the 2×2 matrix formed by these vectors must be zero:

$$\det \begin{pmatrix} x_{j_1} & x_{j_2} \\ y_{j_1} & y_{j_2} \end{pmatrix} = 0 \implies \det \begin{pmatrix} P_1^{(j_1)}(r_1) & P_1^{(j_2)}(r_1) \\ P_0^{(j_1)}(r_1) & P_0^{(j_2)}(r_1) \end{pmatrix} = 0 \quad (5.1)$$

Of course, this observation must be justified rigorously, and this is done explicitly in the proof.

At this point, our goal is to apply Lemma 5.6 to derive a lower bound on the height of $P(x, y)$. To do so, we need to construct a univariate polynomial $Q(x)$ satisfying the conditions of said lemma. The determinant (5.1) provides a natural way to do so; if we replace the evaluation point r_1 in the determinant with a variable x , the expression becomes univariate. Moreover, the identity (5.1) suggests that r_1 may be a root of $Q(x)$ of high multiplicity, which is precisely what is required to invoke Lemma 5.6.

With this construction in place, we can prove the proposition:

Proof. Since $\partial_x^j P(\mathbf{r}) = 0$ for $j = 0, 1, \dots, l-1$, we can write:

$$P_1^{(j)}(r_1)r_2 + P_0^{(j)}(r_1) = 0 \quad \text{for } j = 0, 1, \dots, l-1 \quad (5.2)$$

Claim 1. For any $0 \leq j_1, j_2 \leq l-1$,

$$\det \begin{pmatrix} P_1^{(j_1)}(r_1) & P_1^{(j_2)}(r_1) \\ P_0^{(j_1)}(r_1) & P_0^{(j_2)}(r_1) \end{pmatrix} = 0 \quad (5.3)$$

Proof of the claim 1. Multiplying the determinant by r_2 and expanding the terms, we obtain

$$\begin{aligned} r_2 \cdot \det \begin{pmatrix} P_1^{(j_1)}(r_1) & P_1^{(j_2)}(r_1) \\ P_0^{(j_1)}(r_1) & P_0^{(j_2)}(r_1) \end{pmatrix} &= r_2 [P_1^{(j_1)}(r_1)P_0^{(j_2)}(r_1) - P_1^{(j_2)}(r_1)P_0^{(j_1)}(r_1)] \\ &= \underbrace{[r_2 P_1^{(j_1)}(r_1) + P_0^{(j_1)}(r_1)]}_{=0} P_0^{(j_2)}(r_1) - P_0^{(j_1)}(r_1) P_0^{(j_2)}(r_1) \\ &\quad - \underbrace{[r_2 P_1^{(j_2)}(r_1) + P_0^{(j_2)}(r_1)]}_{=0} P_0^{(j_1)}(r_1) + P_0^{(j_2)}(r_1) P_0^{(j_1)}(r_1) \end{aligned}$$

$$= -P_0^{(j_1)}(r_1)P_0^{(j_2)}(r_1) + P_0^{(j_2)}(r_1)P_0^{(j_1)}(r_1) = 0$$

where we used equation (5.2) to eliminate the first and third terms in the second equality. Hence, the claim holds.

Next, we define the polynomial by replacing the evaluation point r_1 in the determinant (5.3) with a variable x and choosing $j_1 = 0$ and $j_2 = 1$ for simplicity. That is,

$$Q(x) := \det \begin{pmatrix} P_1(x) & P_1'(x) \\ P_0(x) & P_0'(x) \end{pmatrix}.$$

Claim 2. If $l \geq 2$, the polynomial $Q(x)$ has a root at r_1 with multiplicity at least $l - 1$, i.e.

$$Q^{(j)}(r_1) = 0 \quad \text{for } j = 0, 1, \dots, l - 2$$

Proof of the claim 2. In order to find an explicit formula for the derivatives of $Q(x)$, we recall the general Leibniz rule for higher-order derivatives, which states:

$$(fg)^{(n)} = \sum_{i=0}^n \binom{n}{i} f^{(i)} g^{(n-i)}.$$

Since $Q(x)$ is given in terms of a determinant, we cannot directly apply the *Leibniz Rule*. However, we can use the following approach:

$$\begin{aligned} Q^{(j)}(x) &= \left(\det \begin{pmatrix} P_1(x) & P_1'(x) \\ P_0(x) & P_0'(x) \end{pmatrix} \right)^{(j)} = (P_1(x)P_0'(x) - P_0(x)P_1'(x))^{(j)} \\ &= \sum_{i=0}^j \binom{j}{i} \left(P_1^{(i)}(x)P_0^{(j-i+1)}(x) - P_0^{(i)}(x)P_1^{(j-i+1)}(x) \right) \\ &= \sum_{i=0}^j \binom{j}{i} \det \begin{pmatrix} P_1^{(i)}(x) & P_1^{(j-i+1)}(x) \\ P_0^{(i)}(x) & P_0^{(j-i+1)}(x) \end{pmatrix} \end{aligned}$$

By equation (5.3), each determinant in the sum is zero. Hence, $Q^{(j)}(r_1) = 0$ for $j = 0, 1, \dots, l - 2$. Note that this argument does not extend to $j = l - 1$, since the sum then involves $P_1^{(l)}(r_1)$ and $P_0^{(l)}(r_1)$, which are not constrained by our hypothesis. Thus, the claim is proved.

We now establish the majority of the conditions needed to apply Lemma 5.6 to the polynomial $Q(x)$. However, we must first consider whether $Q(x)$ is the zero polynomial. We consider two cases:

Case I: If $Q(x) \neq 0$, then in order to apply the claim above, we need to consider two subcases:

(a) For $l = 1$, since $Q(x)$ is non-zero, we have

$$|P| \geq \max(|P_0|, |P_1|) \geq 1 \geq \frac{1}{2D} = \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right)$$

(b) For $l \geq 2$, our claim shows that $Q^{(j)}(r_1) = 0$ for $j = 0, 1, \dots, l - 2$. Therefore, we can directly apply Lemma 5.6 to obtain $|Q| \geq \|r_1\|^{l-1}$. Writing two polynomials as $P_0(x) = \sum_{i=0}^D a_i x^i$ and $P_1(x) = \sum_{i=0}^D b_i x^i$, we have

$$Q(x) = P_1(x)P_0'(x) - P_0(x)P_1'(x)$$

$$\begin{aligned}
&= \left(\sum_{i=0}^D b_i x^i \right) \left(\sum_{j=1}^D j a_j x^{j-1} \right) - \left(\sum_{i=0}^D a_i x^i \right) \left(\sum_{j=1}^D j b_j x^{j-1} \right) \\
&= \sum_{i=0}^D \sum_{j=1}^D j (b_i a_j - a_i b_j) x^{i+j-1}
\end{aligned}$$

Therefore,

$$|Q| \leq 2D \cdot \max_{0 \leq i, j \leq D} |j(b_i a_j - a_i b_j)| \leq 2D \cdot D \cdot 2|P|^2 = 4D^2|P|^2$$

Combining the upper and lower bounds for $|Q|$ gives

$$\|r_1\|^{l-1} \leq |Q| \leq 4D^2|P|^2 \implies |P| \geq \frac{1}{2D} \|r_1\|^{\frac{l-1}{2}} \geq \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right)$$

Case II: If $Q(x) = 0$, then $P_1(x)P'_0(x) - P_0(x)P'_1(x) = 0$, and we consider two subcases according to whether $P_1(x) = 0$.

- (a) If $P_1 = 0$, then $P_0(x) = P(x, y) \neq 0$ and $P_0^{(j)}(r_1) = 0$ for $j = 0, 1, \dots, l-1$. By Lemma 5.6 again, we have

$$|P| = |P_0| \geq \|r_1\|^l \geq \frac{1}{2D} \|r_1\|^{\frac{l-1}{2}} \geq \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right)$$

- (b) If $P_1 \neq 0$, notice that

$$\left(\frac{P_0}{P_1} \right)' = \frac{P'_0 P_1 - P_0 P'_1}{P_1^2} = \frac{Q}{P_1^2} = 0$$

hence $\frac{P_0}{P_1}$ is constant. Write $P_0 = cP_1$ for some $c \in \mathbb{Q}$, so that $P(x, y) = P_1(x)(y + c) \neq 0$. Condition $\partial_x^j P(\mathbf{r}) = 0$ becomes $P_1^{(j)}(r_1)(r_2 + c) = 0$ for $j = 0, 1, \dots, l-1$.

- (i) If $r_2 + c \neq 0$, then $P_1^{(j)}(r_1) = 0$ for $j = 0, 1, \dots, l-1$. By Lemma 5.6,

$$|P| \geq |P_1| \geq \|r_1\|^l \geq \frac{1}{2D} \|r_1\|^{\frac{l-1}{2}} \geq \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right)$$

- (ii) If $r_2 + c = 0$, then $c = -r_2 = \frac{-p_2}{q_2}$ in reduced form. Then

$$P(x, y) = P_1(x) \left(y - \frac{p_2}{q_2} \right) = \frac{1}{q_2} P_1(x) (q_2 y - p_2) = \frac{c(P_1)}{q_2} \widetilde{P}_1(x) (q_2 y - p_2)$$

where $c(P_1)$ is the content of P_1 and $\widetilde{P}_1(x) \in \mathbb{Z}[x]$ is primitive. Multiplying both sides by q_2 and taking the content of both sides, using Gauss's Lemma 5.4 and Proposition 5.5, we obtain

$$c(q_2 \cdot P) = c(c(P_1) \cdot \widetilde{P}_1 \cdot (q_2 y - p_2)) \implies |q_2| \cdot c(P) = |c(P_1)| \cdot 1 \cdot 1 = |c(P_1)|$$

Thus, we have $\frac{c(P_1)}{q_2} = \pm 1$ and hence $P(x, y) = \pm \widetilde{P}_1(x) (q_2 y - p_2)$. Therefore, we conclude that

$$|P| \geq \max(|p_2|, |q_2|) = \|r_2\| \geq \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right)$$

Finally, combining all the cases above, we have proved the proposition. $\square$

5.3 Thue's Theorem

Having established the necessary lemmas and propositions, we can now prove Thue's theorem. As discussed, we cannot guarantee that a given approximation is non-zero when evaluated at the polynomial. To resolve this, we consider the derivative $\partial_x^j P(x, y)$, where j is bounded and chosen such that $\partial_x^j P(r_1, r_2) \neq 0$, for some rational approximation (r_1, r_2) . To bound j , we shall relate the multiplicity of our rational approximation (r_1, r_2) , denoted l , to the multiplicity at (β, β) , m . We have already established an upper bound $|P| < B(\beta, \varepsilon)^m$, and a corresponding lower bound of the form $|P| > B(\beta, \varepsilon)^{f(l)}$ would yield a relationship between l and m . We can obtain such a lower bound from Proposition 5.7, assuming that a given rational approximation (r_1, r_2) is such that $\frac{1}{2D} \|r_1\| \leq \|r_2\|$.

Theorem 5.8 (Thue's Theorem [2, Lecture 25]). *Any algebraic number β of degree $d \geq 2$ is not approximable by rationals to any order greater than $\frac{d}{2} + 1$. More precisely, for any $\delta > 0$ and any constant $K = K(\beta) > 0$, the inequality*

$$\left| \beta - \frac{p}{q} \right| \leq \frac{K(\beta)}{q^{d/2+1+\delta}}$$

holds for only finitely many rationals $\frac{p}{q}$ in reduced form.

Proof. Fix $0 < \varepsilon < \frac{1}{4}$, $\delta > 0$ and suppose for some $\gamma > 0$ we have that

$$\left| \beta - \frac{p}{q} \right| < \frac{K(\beta)}{q^\gamma},$$

holds for some constant $K(\beta) > 0$ and infinitely many rationals p/q . Since there are infinitely many rationals, we may choose two distinct reduced rationals $r_1 = p_1/q_1$ and $r_2 = p_2/q_2$ with $r_1, r_2 \in I_\beta := [\beta - 1, \beta + 1]$ and $q_1, q_2 > 0$ such that

$$|\beta - r_i| < \frac{K(\beta)}{q_i^\gamma} \quad \text{for } i = 1, 2. \quad (5.4)$$

We further choose r_1 and r_2 such that:

- (a) $q_1 \geq A(\beta, \varepsilon)$, for a constant $A(\beta, \varepsilon) > 0$ to be fixed later;
- (b) $q_2 \geq q_1^{1/\varepsilon}$.

With these choices, we now fix (r_1, r_2) . Since the intervals $[q_1^{(1-\varepsilon)m-1}, q_1^{(1-\varepsilon)(m+1)-1}]$ for $m = 1, 2, \dots$ cover $[1, \infty)$, we may choose $m \in \mathbb{Z}^+$ such that for q_1, q_2 and ε we have,

$$q_1^{(1-\varepsilon)m-1} \leq q_2 \leq q_1^{(1-\varepsilon)(m+1)-1}. \quad (5.5)$$

Applying Proposition 5.2 to m , β , and ε , we obtain a non-zero polynomial

$$P(x, y) = P_1(x)y + P_0(x) \in \mathbb{Z}[x, y]$$

satisfying:

- (c) $\partial_x^j P(\beta, \beta) = 0$ for $j = 0, 1, \dots, m-1$;
- (d) $|P| \leq B(\beta, \varepsilon)^m$;
- (e) $\deg(P_0), \deg(P_1) \leq (1 + \varepsilon)\frac{d}{2}m$.

Let $D = \max\{\deg(P_0), \deg(P_1)\}$. Since $D \leq (1 + \varepsilon)\frac{d}{2}m$ and $(1 + \varepsilon)\frac{d}{2}m$ grows linearly in m while $B(\beta, \varepsilon)^m$ grows exponentially, there exists a smallest positive integer $m_0 = m_0(\beta, \varepsilon)$ such that $2D \leq (1 + \varepsilon)dm \leq B(\beta, \varepsilon)^m$ for all $m \geq m_0$. Define $D_0 = D_0(\beta, \varepsilon) :=$

$\lfloor (1+\varepsilon)\frac{d}{2}m_0 \rfloor$, so that $D \leq D_0$ for all $m \leq m_0$. To apply Proposition 5.7, we may assume that P is primitive.

Claim 1. $q_2 > B(\beta, \varepsilon)^m$ whenever $A(\beta, \varepsilon) > [2D_0B(\beta, \varepsilon)]^{8/\varepsilon}$.

Proof of Claim 1. By using inequalities (5.5) and (a) we have, $q_2 \geq q_1^{(1-\varepsilon)m-1} \geq A(\beta, \varepsilon)^{(1-\varepsilon)m-1}$. We shall now impose our first condition on $A(\beta, \varepsilon)$. Let:

$$A(\beta, \varepsilon) > [2D_0B(\beta, \varepsilon)]^{\frac{8}{\varepsilon}}. \quad (5.6)$$

Note that since $0 < \varepsilon < \frac{1}{4}$ and $D_0 \geq 1$, we have $A(\beta, \varepsilon) > [2D_0B(\beta, \varepsilon)]^{\frac{8}{\varepsilon}} > B(\beta, \varepsilon)^4$, which implies that

$$\begin{aligned} \frac{\ln(A(\beta, \varepsilon))}{\ln(B(\beta, \varepsilon))} &> 4 = \frac{1}{1 - (\frac{1}{4} + \frac{1}{2})} \geq \frac{1}{1 - (\varepsilon + \frac{1}{m})} = \frac{m}{(1-\varepsilon)m-1} \\ \implies \ln(A(\beta, \varepsilon))((1-\varepsilon)m-1) &> \ln(B(\beta, \varepsilon))m \\ \implies A(\beta, \varepsilon)^{(1-\varepsilon)m-1} &> B(\beta, \varepsilon)^m. \end{aligned}$$

Thus, it follows that $q_2 > B(\beta, \varepsilon)^m$, and this completes the proof of the claim.

Let $l \geq 1$ be the order of vanishing of $P(x, y)$ at the point (r_1, r_2) with respect to x , i.e. l is the largest integer such that $\partial_x^j P(r_1, r_2) = 0$ for $j = 0, 1, \dots, l-1$ (so that $\partial_x^l P(r_1, r_2) \neq 0$). Then by Proposition 5.7, we have

$$|P| \geq \min \left(\frac{1}{2D} \|r_1\|^{\frac{l-1}{2}}, \|r_2\| \right) \geq \min \left(\frac{1}{2D} q_1^{\frac{l-1}{2}}, q_2 \right).$$

Suppose $l \geq 2$; we shall treat the case $l = 1$ afterward. If $\min(\frac{1}{2D} q_1^{\frac{l-1}{2}}, q_2) = q_2$, then by the inequality (d) and Claim 1, we have

$$B(\beta, \varepsilon)^m \geq |P| \geq \min \left(\frac{1}{2D} q_1^{\frac{l-1}{2}}, q_2 \right) = q_2 > B(\beta, \varepsilon)^m.$$

This is a contradiction. Thus we must have $\min(\frac{1}{2D} q_1^{\frac{l-1}{2}}, q_2) = \frac{1}{2D} q_1^{\frac{l-1}{2}}$. Combining this with the inequality (a), we have:

$$B(\beta, \varepsilon)^m \geq |P| \geq \min \left(\frac{1}{2D} q_1^{\frac{l-1}{2}}, q_2 \right) = \frac{1}{2D} q_1^{\frac{l-1}{2}} \geq \frac{1}{2D} A(\beta, \varepsilon)^{\frac{l-1}{2}}.$$

We shall now bound l in terms of m :

Claim 2. $l < \varepsilon m$.

Proof of Claim 2. From the chain of inequalities above (line ((d)) and Proposition 5.7), we have

$$B(\beta, \varepsilon)^m \geq \frac{1}{2D} A(\beta, \varepsilon)^{\frac{l-1}{2}}. \quad (5.7)$$

We consider two cases according to the size of m .

Case $m \leq m_0$: Since $D \leq D_0$ and $A(\beta, \varepsilon) > [2D_0B(\beta, \varepsilon)]^{8/\varepsilon} \geq [2DB(\beta, \varepsilon)]^{8/\varepsilon}$, we have

$$\frac{\ln A(\beta, \varepsilon)}{\ln(2DB(\beta, \varepsilon))} > \frac{8}{\varepsilon} \geq \frac{4}{\varepsilon} \frac{l}{l-1},$$

which gives $\frac{1}{2D} A(\beta, \varepsilon)^{(l-1)/2} > B(\beta, \varepsilon)^{2l/\varepsilon}$. Combined with (5.7), this yields $B^m \geq B^{2l/\varepsilon}$, hence $l \leq \varepsilon m/2 < \varepsilon m$.

Case $m > m_0$: By the definition of m_0 , we have $\ln(2D) \leq m \ln B$. Combining this with (5.7):

$$2m \ln B \geq m \ln B + \ln(2D) \geq \frac{l-1}{2} \ln A > \frac{l-1}{2} \cdot \frac{8}{\varepsilon} \ln B = \frac{4(l-1)}{\varepsilon} \ln B,$$

where the last inequality uses $A > [2D_0B]^{8/\varepsilon} > B^{8/\varepsilon}$. Dividing by $\ln B > 0$, we obtain $2m > 4(l-1)/\varepsilon$, which gives $l < 1 + \varepsilon m/2$. Since $l \geq 2$ implies $\varepsilon m > 2$, we have $1 + \varepsilon m/2 < \varepsilon m$, and hence $l < \varepsilon m$.

In both cases, $l < \varepsilon m$. This completes the proof of the claim.

We can use this relationship to find a $j \leq \varepsilon m$ such that $\partial_x^j P(r_1, r_2) \neq 0$.

Claim 3. $\exists j \leq \varepsilon m$ such that $\partial_x^j P(r_1, r_2) \neq 0$.

Proof of Claim 3. By definition, l is the vanishing order of P at (r_1, r_2) with respect to x , so $\partial_x^l P(r_1, r_2) \neq 0$. By Claim 2, $l < \varepsilon m$, and since l is a positive integer, $l \leq \lfloor \varepsilon m \rfloor \leq \varepsilon m$. Hence $j = l$ satisfies the claim. Therefore Claim 3 is proved.

Now we consider the remaining case $l = 1$. If $l = 1$, set $j = 1$. Then $\partial_x P(r_1, r_2) \neq 0$. Combining conditions (5.5) and (b), we obtain

$$q_1^{\frac{1}{\varepsilon}} \leq q_2 \leq q_1^{(1-\varepsilon)(m+1)-1} \implies (1-\varepsilon)(m+1) - 1 \geq \frac{1}{\varepsilon} \implies m \geq \frac{1}{\varepsilon}. \quad (5.8)$$

Hence $j = 1 \leq \varepsilon m$.

So in all cases, there exists some $j \leq \varepsilon m$ such that $\partial_x^j P(r_1, r_2) \neq 0$. Let

$$\tilde{P}(x, y) = \frac{1}{j!} \partial_x^j P(x, y) = \tilde{P}_1(x)y + \tilde{P}_0(x) \in \mathbb{Z}[x, y]$$

where $\tilde{P}_1(x) = \frac{1}{j!} P_1^{(j)}(x)$ and $\tilde{P}_0(x) = \frac{1}{j!} P_0^{(j)}(x)$. Then we have

(A) $\tilde{P}(r_1, r_2) \neq 0$

(B) $\partial_x^k \tilde{P}(\beta, \beta) = \frac{1}{j!} \partial_x^{k+j} P(\beta, \beta) = 0$ for $k = 0, 1, \dots, \lfloor (1-\varepsilon)m \rfloor - 1$

Note that when $k \leq \lfloor (1-\varepsilon)m \rfloor - 1 \leq (1-\varepsilon)m - 1$, we have $k+j \leq (1-\varepsilon)m - 1 + \varepsilon m = m - 1$, thus by condition (c), we have $\partial_x^{k+j} P(\beta, \beta) = 0$.

(C) $\deg(\tilde{P}_0), \deg(\tilde{P}_1) \leq (1+\varepsilon)\frac{d}{2}m$

Let $\tilde{D} = \max(\deg(\tilde{P}_0), \deg(\tilde{P}_1)) \leq (1+\varepsilon)\frac{d}{2}m$.

(D) $|\tilde{P}| \leq \tilde{B}(\beta, \varepsilon)^m$ for some constant $\tilde{B}(\beta, \varepsilon) > 0$ depending only on β and ε .

The last property (D) holds since

$$\begin{aligned} \tilde{P} &= \frac{1}{j!} \partial_x^j P(x, y) = \sum_{i=j}^D \binom{i}{j} b_i x^{i-j} y + \sum_{i=j}^D \binom{i}{j} a_i x^{i-j} \\ \implies |\tilde{P}| &\leq \max_{j \leq i \leq D} \left\{ \left| \binom{i}{j} a_i \right|, \left| \binom{i}{j} b_i \right| \right\} \leq \binom{D}{j} |P| \leq 2^D |P| \\ &\leq 2^{(1+\varepsilon)\frac{d}{2}m} B(\beta, \varepsilon)^m = \left(2^{(1+\varepsilon)\frac{d}{2}} B(\beta, \varepsilon) \right)^m =: \tilde{B}(\beta, \varepsilon)^m \end{aligned}$$

We can now derive a lower bound for $|\tilde{P}(r_1, r_2)|$ as in the proof of Liouville's theorem, since $\tilde{P}(r_1, r_2) \neq 0$.

$$0 \neq |\tilde{P}(r_1, r_2)| = |\tilde{P}_1(r_1)r_2 + \tilde{P}_0(r_1)| = \frac{|\text{some non-zero integer}|}{q_1^{\tilde{D}} q_2} \geq \frac{1}{q_1^{\tilde{D}} q_2}.$$

We can also derive an upper bound for $|\tilde{P}(r_1, r_2)|$. Notice that

$$\begin{aligned}\tilde{P}(r_1, r_2) &= \underbrace{\tilde{P}(r_1, r_2) - \tilde{P}(r_1, \beta)}_{=\tilde{P}_1(r_1)(r_2 - \beta)} + \underbrace{\tilde{P}(r_1, \beta)}_{\text{expand via Taylor's theorem}} \\ &= \tilde{P}_1(r_1)(r_2 - \beta) + \sum_{j=0}^{\tilde{D}} \frac{\partial_x^j \tilde{P}(\beta, \beta)}{j!} (r_1 - \beta)^j \\ &= \tilde{P}_1(r_1)(r_2 - \beta) + \sum_{j=J_{m,\varepsilon}}^{\tilde{D}} \frac{\partial_x^j \tilde{P}(\beta, \beta)}{j!} (r_1 - \beta)^j \quad \text{by (B) and } J_{m,\varepsilon} = \lfloor (1 - \varepsilon)m \rfloor\end{aligned}$$

Recall that $r_1, r_2 \in I_\beta = [\beta - 1, \beta + 1]$, so $|r_1|, |r_2| \leq |\beta| + 1$, and using (5.4) we have

$$\begin{aligned}|\tilde{P}(r_1, r_2)| &\leq \underbrace{|\tilde{P}_1(r_1)|}_{\leq (1+|\beta|)^{\tilde{D}}(\tilde{D}+1)\tilde{B}(\beta, \varepsilon)^m} \underbrace{|r_2 - \beta|}_{\leq \frac{K(\beta)}{q_2}} + \underbrace{\sum_{j=J_{m,\varepsilon}}^{\tilde{D}} \left| \frac{\partial_x^j \tilde{P}(\beta, \beta)}{j!} \right|}_{\leq (\tilde{D}+1)C(\beta, \varepsilon)^m} \underbrace{|r_1 - \beta|^j}_{\leq \frac{K(\beta)^{\tilde{D}}}{q_1^{J_{m,\varepsilon}\gamma}}}\end{aligned}$$

and hence

$$|\tilde{P}(r_1, r_2)| \leq (1 + |\beta|)^{\tilde{D}}(\tilde{D} + 1)\tilde{B}(\beta, \varepsilon)^m K(\beta) \frac{1}{q_2^\gamma} + (\tilde{D} + 1)C(\beta, \varepsilon)^m K(\beta)^{\tilde{D}} \frac{1}{q_1^{J_{m,\varepsilon}\gamma}}.$$

where $C(\beta, \varepsilon) > 0$ is defined in the same manner as $\tilde{B}(\beta, \varepsilon)$ in property (D). Thus, by introducing a new constant $E(\beta, \varepsilon) > 0$ depending only on β and ε and not on m, q_1, q_2 , we can simplify the above to

$$|\tilde{P}(r_1, r_2)| \leq E(\beta, \varepsilon)^m \left(\frac{1}{q_2^\gamma} + \frac{1}{q_1^{J_{m,\varepsilon}\gamma}} \right) \leq E(\beta, \varepsilon)^m \left[\frac{1}{q_2^\gamma} + \frac{1}{q_1^{[(1-\varepsilon)m-1]\gamma}} \right].$$

This provides us with an upper bound for $|\tilde{P}(r_1, r_2)|$. Replacing q_2^γ by $q_1^{[(1-\varepsilon)m-1]\gamma}$ via (5.5), we obtain

$$|\tilde{P}(r_1, r_2)| \leq 2E(\beta, \varepsilon)^m \frac{1}{q_1^{[(1-\varepsilon)m-1]\gamma}}. \quad (5.9)$$

Combining the lower bound and upper bound for $|\tilde{P}(r_1, r_2)|$, we have

$$\begin{aligned}\frac{1}{q_1^{\tilde{D}} q_2} &\leq |\tilde{P}(r_1, r_2)| \leq 2E(\beta, \varepsilon)^m \frac{1}{q_1^{[(1-\varepsilon)m-1]\gamma}} \\ \implies \frac{1}{q_1^{(1+\varepsilon)\frac{d}{2}m}} \frac{1}{q_2} &\leq 2E(\beta, \varepsilon)^m \frac{1}{q_1^{[(1-\varepsilon)m-1]\gamma}} \quad \text{by property (C)} \\ \implies q_1^{[(1-\varepsilon)m-1]\gamma} &\leq 2E(\beta, \varepsilon)^m q_1^{(1+\varepsilon)\frac{d}{2}m} q_1^{(1-\varepsilon)(m+1)-1} \quad \text{by inequality (5.5)}\end{aligned}$$

Claim 4. $q_1^{(1-2\varepsilon)m\gamma} \leq q_1^{[(1-\varepsilon)m-1]\gamma}$.

Proof of Claim 4. Note that by inequality (5.8) we have $\varepsilon m \geq 1$. Hence

$$[(1 - \varepsilon)m - 1] - (1 - 2\varepsilon)m = -1 + \varepsilon m \geq 0 \implies [(1 - \varepsilon)m - 1] \geq (1 - 2\varepsilon)m.$$

Since $\gamma > 0$, it follows that $q_1^{(1-2\varepsilon)m\gamma} \leq q_1^{[(1-\varepsilon)m-1]\gamma}$. This completes the proof of the claim.

Applying Claim 4, we can replace $q_1^{[(1-\varepsilon)m-1]\gamma}$ with $q_1^{(1-2\varepsilon)m\gamma}$ and obtain:

$$\begin{aligned} q_1^{(1-2\varepsilon)m\gamma} &\leq q_1^{[(1-\varepsilon)m-1]\gamma} \leq 2E(\beta, \varepsilon)^m q_1^{(1+\varepsilon)\frac{d}{2}m} q_1^{(1-\varepsilon)(m+1)-1} \\ \implies q_1^{(1-2\varepsilon)\gamma} &\leq 2E(\beta, \varepsilon) q_1^{(1+\varepsilon)\frac{d}{2}+(1-\varepsilon)(1+\frac{1}{m})} \leq 2E(\beta, \varepsilon) q_1^{(1+\varepsilon)\frac{d}{2}+(1-\varepsilon)(1+\varepsilon)} \text{ since } m \geq \frac{1}{\varepsilon} \\ \implies (1-2\varepsilon)\gamma &\leq \frac{\ln[2E(\beta, \varepsilon)]}{\ln[q_1]} + (1+\varepsilon)\frac{d}{2} + 1 - \varepsilon^2. \end{aligned}$$

Using condition (a), we impose a second and final condition on $A(\beta, \varepsilon)$. Let

$$A(\beta, \varepsilon) \geq e^{\ln[2E(\beta, \varepsilon)]\frac{2}{\delta}} \quad (5.10)$$

To satisfy (5.6) and (5.10) we fix $A(\beta, \varepsilon)$ with

$$A(\beta, \varepsilon) = \max \left\{ [2D_0(\beta, \varepsilon) B(\beta, \varepsilon)]^{\frac{8}{\varepsilon}}, e^{\frac{2}{\delta} \ln[2E(\beta, \varepsilon)]} \right\} + 1.$$

Thus, we have $\frac{\ln[2E(\beta, \varepsilon)]}{\ln[q_1]} \leq \frac{\delta}{2}$ and hence

$$(1-2\varepsilon)\gamma \leq \frac{\delta}{2} + (1+\varepsilon)\frac{d}{2} + 1 - \varepsilon^2.$$

Since $\varepsilon > 0$ can be chosen to be arbitrarily small, we let $\varepsilon \rightarrow 0$ and obtain

$$\gamma \leq \frac{d}{2} + 1 + \frac{\delta}{2} < \frac{d}{2} + 1 + \delta.$$

This means that if there are infinitely many rationals satisfying the inequality (5.4), then the value of γ is less than $\frac{d}{2} + 1 + \delta$. Therefore, we conclude that for any $\delta > 0$ and any constant $K = K(\beta) > 0$, the inequality

$$\left| \beta - \frac{p}{q} \right| \leq \frac{K}{q^{d/2+1+\delta}}$$

holds for only finitely many rationals $\frac{p}{q}$ in reduced form. $\square$

The proof of Thue's theorem now completes the argument unfinished in Chapter 4. In Chapter 4, after reducing the equation defined by a binary homogeneous form to a problem of rational approximation, Liouville's theorem ruled out all cases except the possibility that a real algebraic number of degree at least 3 admits infinitely many rational approximations with the required exponent. Thue's theorem removes this remaining possibility. Hence, apart from the two explicit exceptional families identified in Theorem 4.12, the equation $F(x, y) = m$ admits only finitely many integer solutions.

This also marks a natural endpoint for the method of rational approximation developed so far. Both Liouville's theorem and Thue's theorem are results establishing finiteness, and so they are best suited to Diophantine equations whose integer solutions are finite in number. However, not all Diophantine equations fall into this setting. A classical example is Pell's equation, which admits infinitely many integer solutions. To understand such equations, one must move beyond approximation bounds and instead study the arithmetic structure of the underlying number field. This is the purpose of the next chapter, where we turn to Dirichlet's Unit Theorem and see how the unit group of the ring of integers gives rise to infinite families of solutions.

Chapter 6

Pell Equations and Dirichlet's Unit Theorem

In Chapter 4, we used Liouville's Theorem to classify the integer solutions of binary homogeneous polynomials, thereby establishing a link between rational approximation and Diophantine equations. For one of the main classes of polynomials, however, Liouville's Theorem was not strong enough to give us a definitive solution set. This led us to Thue's Theorem in Chapter 5.

Consider the binary form

$$x^2 - Dy^2 = 1, \quad (6.1)$$

where D is a positive square-free integer. Diophantine equations of this form are often referred to as *Pell's equation*. The scrutiny of Pell's equation dates back to 400 BC, and the classification of its solutions reveals a deep connection between the algebraic structure of the field $\mathbb{Q}(\sqrt{D})$ and rational approximations of $\sqrt{D}$.

A particularly fruitful strategy is to reinterpret the solution set in terms of units in a quadratic ring. Let $R := \mathbb{Z}[\sqrt{D}]$. A theorem of Dirichlet tells us that the group of units $U := R^\times$ is isomorphic to the group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}$. This is equivalent to the statement

$$U = \{\pm \varepsilon^k : k \in \mathbb{Z}\},$$

for some $\varepsilon \in U \setminus \{\pm 1\}$. Such a unit is called a *fundamental unit* if $U = \{\pm \varepsilon^k : k \in \mathbb{Z}\}$. Now, how does this relate to the above equation? Well, if one can argue that

$$\alpha = a + b\sqrt{D} \in U \iff a^2 - Db^2 = \pm 1, \quad (6.2)$$

then the aforementioned theorem of Dirichlet tells us that the set $\{\pm \varepsilon^k : k \in \mathbb{Z}\}$ is precisely the unit group U , and hence describes exactly the elements $\alpha = a + b\sqrt{D}$ satisfying $a^2 - Db^2 = \pm 1$. This will not only allow us to establish the cardinality of the solution set of (6.1), but also to characterise the solutions completely.

Let us start by analysing the structure of the ring R . We can regard it as a subring of the field $\mathbb{Q}(\sqrt{D})$. Since $\sqrt{D}$ is algebraic over $\mathbb{Q}$ with minimal polynomial $x^2 - D \in \mathbb{Q}[x]$, it follows that $\mathbb{Q}(\sqrt{D}) = \{a + b\sqrt{D} : a, b \in \mathbb{Q}\}$ is an extension over $\mathbb{Q}$ of degree 2. This brings us to the definition of a *number field*:

Definition 6.1. A *number field* is a finite field extension over $\mathbb{Q}$.

It turns out that we can recover (6.2) by considering the *norm* and the *trace* of the

elements of $\mathbb{Q}(\sqrt{D})$, which are defined as follows for a general extension L/K of number fields:

Definition 6.2. Let L/K be an extension of number fields, and let $\alpha \in L$. Viewing L as a vector space over K , consider the linear map $m_\alpha : L \rightarrow L$, defined by $m_\alpha(\ell) := \alpha\ell$. The *norm* of α is defined as:

$$N_{L/K}(\alpha) := \det m_\alpha,$$

and the *trace* of α is defined as follows:

$$\text{tr}_{L/K}(\alpha) := \text{tr } m_\alpha.$$

Note that linearity of $\text{tr}_{L/K}$ and the fact that $N_{L/K}$ is multiplicative follow as a result of the properties of the determinant and the trace.

Let us return to the extension $\mathbb{Q}(\sqrt{D})/\mathbb{Q}$ of number fields. In this case, the norm and trace of an element $\alpha = a + b\sqrt{D} \in \mathbb{Q}(\sqrt{D})$ are given by:

$$N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(\alpha) = \det \begin{pmatrix} a & bD \\ b & a \end{pmatrix} = a^2 - Db^2, \quad \text{tr}_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(\alpha) = 2a.$$

We now have an explicit link to (6.2). For $\alpha \in R$, we can use the shorthand notation $N(\alpha)$ for $N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(\alpha)$. Note that the above implies that $N(\alpha)$ is an integer. Furthermore, if $b \neq 0$,

$$a^2 - Db^2 = 0 \implies Db^2 = a^2 \implies \sqrt{D} = a/b \in \mathbb{Q},$$

a contradiction. If $b = 0$, then $a^2 - Db^2 = 0 \implies a = 0$. Hence, the norm takes non-zero integer values for non-zero elements of R .

If we can establish that $|N(\alpha)| = 1$ if and only if $\alpha \in U$, then we will have shown that (6.2) holds. We therefore prove the following proposition:

Proposition 6.3 ([4]). *Let $\alpha \in R$. Then $\alpha \in U \iff |N(\alpha)| = 1$.*

Before we prove this result, we recall that the conjugate $\tilde{\alpha}$ of $\alpha = a + b\sqrt{D} \in \mathbb{Q}(\sqrt{D})$ is given by $\tilde{\alpha} := a - b\sqrt{D}$.

Proof. By our previous discussion, we have $N(\alpha) = \alpha\tilde{\alpha}$. Hence, if $\alpha \in R$ satisfies $N(\alpha) = \pm 1$, then α is a unit with inverse $\tilde{\alpha}$ (if $N(\alpha) = 1$) or $-\tilde{\alpha}$ (if $N(\alpha) = -1$). Conversely, if α is a unit, then $\alpha\beta = 1$ for some $\beta \in R$. As the norm is multiplicative, we then have $N(\alpha)N(\beta) = N(\alpha\beta) = 1$. Since $N(\alpha)$ and $N(\beta)$ are integers, this forces $N(\alpha) = \pm 1$. $\square$

The following lemma will also be useful in analysing the structure of the ring:

Lemma 6.4 ([4]). *For $\alpha = a + b\sqrt{D} \in R$ such that $b \neq 0$, the minimal polynomial f_α of α over $\mathbb{Q}$ is given by $f_\alpha(x) := x^2 - \text{tr}(\alpha)x + N(\alpha)$.*

Proof. Let $\alpha \in R$ be of the above form. Since $b \neq 0$, then $\alpha \notin \mathbb{Q}$. Hence, the degree of the minimal polynomial of α is at least 2. Furthermore,

$$\begin{aligned} f_\alpha(x) &= x^2 - \text{tr}(\alpha)x + N(\alpha) = x^2 - 2ax + (a^2 - Db^2) = x^2 - (\alpha + \tilde{\alpha})x + \alpha\tilde{\alpha} \\ &= (x - \alpha)(x - \tilde{\alpha}), \end{aligned}$$

so f_α annihilates α . It follows that f_α is the minimal polynomial of α over $\mathbb{Q}$. $\square$

Let us now formalise (and prove) the aforementioned theorem of Dirichlet:

Theorem 6.5 (Dirichlet's Unit Theorem for Subrings, [4]). *We have*

$$U = \{\pm \varepsilon^k : k \in \mathbb{Z}\}$$

for some fundamental unit $\varepsilon \notin \{1, -1\}$.

Proof. Let us first give an outline of the proof structure. Consider the following statements:

1. There is an additive group homomorphism $L : U \rightarrow \mathbb{R}$ whose kernel is $\{\pm 1\}$.
2. An additive subgroup G of $\mathbb{R}$ is either dense, or discrete and cyclic. Furthermore, if G is discrete and non-trivial, then $G = a\mathbb{Z}$, where $a = \inf(G \cap \mathbb{R}^+) > 0$.
3. $L(U)$ is discrete and cyclic.
4. There exists a fundamental unit $\varepsilon \in U$.

If (4) holds, then $L(U)$ is non-trivial, since $\varepsilon \notin \{\pm 1\}$ and hence $|\varepsilon| \neq 1$. If (2) and (3) hold, then $L(U) = a\mathbb{Z}$ for some $a > 0$. Hence, there exists some $\varepsilon \in U$ such that $L(\varepsilon) = a$. Such an ε must be a fundamental unit, as $\varepsilon \in \ker L$ otherwise. Let $u \in U$. Since L is additive, we have

$$L(u) = ak = kL(\varepsilon) = L(\varepsilon^k)$$

which implies

$$L(u/\varepsilon^k) = L(u) - L(\varepsilon^k) = 0 \implies |u/\varepsilon^k| = 1 \implies u = \pm \varepsilon^k,$$

for some $k \in \mathbb{Z}$. The result then follows, and the map $\varphi : U \rightarrow \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}$ defined by

$$\varphi(u) = \begin{cases} (0, k) & \text{if } u = \varepsilon^k \text{ for some } k \in \mathbb{Z} \\ (1, k) & \text{if } u = -\varepsilon^k \text{ for some } k \in \mathbb{Z} \end{cases}$$

is an isomorphism. Hence, we proceed by proving the above statements:

Lemma 6.6 ([4]). *Define $L : U \rightarrow \mathbb{R}$ by $L(u) = \log(|u|)$ ¹. Then L is an additive group homomorphism whose kernel is $\{\pm 1\}$.*

Proof. Firstly, we have

$$L(u_1 u_2) = \log(|u_1 u_2|) = \log(|u_1| |u_2|) = \log(|u_1|) + \log(|u_2|) = L(u_1) + L(u_2),$$

for any $u_1, u_2 \in U$. Hence, L is an additive group homomorphism. Furthermore,

$$L(u) = \log(|u|) = 0 \iff |u| = 1 \iff u \in \{1, -1\}.$$

Hence, $\ker L = \{\pm 1\}$. □

Lemma 6.7 ([4]). *An additive subgroup G of $\mathbb{R}$ is either discrete and cyclic, or dense. Furthermore, if G is discrete and non-trivial, then $G = a\mathbb{Z}$, where $a = \inf(G \cap \mathbb{R}^+) > 0$.*

Proof. We begin by noting that the aforementioned properties are mutually exclusive. By definition, a discrete subgroup of $\mathbb{R}$ cannot possibly be dense. Hence, proving that one property holds rules out the other entirely.

If G is trivial, then G is discrete and cyclic.

¹Here, $|u|$ technically refers to $|\iota(u)|$, where ι denotes the inclusion $\iota : \mathbb{Q}(\sqrt{D}) \hookrightarrow \mathbb{C}$. Understanding this nuance will be important later in the chapter.

Now, let G be non-trivial. Since G is an additive subgroup of $\mathbb{R}$, then $x \in G \iff -x \in G$. Now, as G is non-trivial, then there exists a non-zero element of G . Hence, by the above, there exists a strictly positive element of G . Thus, we have $G \cap \mathbb{R}^+ \neq \emptyset$. Furthermore, $G \cap \mathbb{R}^+$ is bounded below by zero. Hence, let $a = \inf(G \cap \mathbb{R}^+)$.

Suppose now that $a = 0$. Let $x \in \mathbb{R}$ and $\varepsilon > 0$ be given. Since $a = \inf(G \cap \mathbb{R}^+)$, it follows that there exists some $g \in G \cap \mathbb{R}^+$ such that $0 < g < \varepsilon$. Since $g \in G \cap \mathbb{R}^+$, this simplifies to $0 < g < \varepsilon$. Note that

$$y = g \cdot \lfloor x/g \rfloor \in G.$$

Furthermore, since $x - g \cdot \lfloor x/g \rfloor \geq x - g(x/g) = 0$, then

$$|x - y| = |x - g \cdot \lfloor x/g \rfloor| = x - g \cdot \lfloor x/g \rfloor = g(x/g - \lfloor x/g \rfloor) \leq g < \varepsilon.$$

Since $x \in \mathbb{R}$ and $\varepsilon > 0$ were taken to be arbitrary, it follows that G is dense in $\mathbb{R}$.

Now, suppose that $a > 0$. Suppose that $a \notin G$. Then, for all $\varepsilon > 0$, there exists $g \in G$ such that $a \leq g < a + \varepsilon$. Taking $\varepsilon = a$, we see that there exists $g_2 \in G$ such that $a \leq g_2 < 2a$. Note that, as $a \notin G$, then $g_2 - a \neq 0$, and thus we can simplify this to $a < g_2 < 2a$. Taking $\varepsilon = g_2 - a$, we see that there exists $g_1 \in G$ such that $a \leq g_1 < g_2$. By the same argument as before, this can again be simplified to $a < g_1 < g_2$. Thus,

$$0 < g_2 - g_1 < 2a - g_1 < 2a - a = a,$$

a contradiction (as $a = \inf(G \cap \mathbb{R}^+)$, and $g_2 - g_1 \in G$). Thus, we must have $a \in G$. Now, let $g \in G$. Then

$$0 = g - a(g/a) \leq g - a \cdot \lfloor g/a \rfloor = a(g/a - \lfloor g/a \rfloor) < a.$$

Since $a = \inf(G \cap \mathbb{R}^+)$, this forces $g - a \cdot \lfloor g/a \rfloor = 0$. That is, $g = a \cdot \lfloor g/a \rfloor \in a\mathbb{Z}$. Since $g \in G$ was taken to be arbitrary, it follows that $G = a\mathbb{Z}$. $\square$

Lemma 6.8 ([4]). *$L(U)$ is discrete and cyclic.*

Proof. Consider the set $S := (-1, 1) \cap L(U)$. By definition of L ,

$$L^{-1}(S) = \{u \in U : \log|u| \in (-1, 1)\} = \{u \in U : e^{-1} \leq |u| \leq e\}.$$

Recall that, for $\alpha := a + b\sqrt{D} \in U$ with $b \neq 0$, the minimal polynomial f_α of α over $\mathbb{Q}$ is

$$f_\alpha(x) := x^2 - \text{tr}(\alpha)x + N(\alpha) = x^2 - (\alpha + \tilde{\alpha})x + N(\alpha).$$

Note that the only elements in U that are not of this form are 1 and -1 , and $L(1) = L(-1) = 0$. For $u \in L^{-1}(S)$, we have $|u\tilde{u}| = |N(u)| = 1$, and hence

$$|u + \tilde{u}| \leq |u| + |\tilde{u}| = |u| + \frac{1}{|u|} \leq e + 1/(e^{-1}) = 2e, \quad |N(u)| = 1.$$

Suppose $f \in \mathbb{Z}[x]$ is a polynomial of the form $x^2 + px + q$, where $|p| \leq 2e$ and $|q| = 1$. Then, we have

$$p \in (-2e, 2e) \cap \mathbb{Z}, \quad q \in \{-1, 1\}.$$

Hence, we have only finitely many choices for p and at most 2 choices for q . There can thus be only finitely many such polynomials. In the field $\mathbb{Q}(\sqrt{D})$, two elements have the same minimal polynomial if and only if they are conjugate. This therefore applies to elements of U , and it follows that $\#L^{-1}(S) \setminus \{\pm 1\}$ must also be finite. Now, define $b \in \mathbb{R}$

by $b := \min\{|L(u)| : u \in L^{-1}(S) \setminus \{\pm 1\}\}$. Then, $(0, b) \cap L(U) = \emptyset$. Thus, $L(U)$ is not dense in $\mathbb{R}$, and it is hence discrete and cyclic. $\square$

Lemma 6.9 ([4]). *There exists a fundamental unit $\varepsilon \in U$.*

Proof. Let $A \subset R$ be defined by

$$A := \{\alpha = a + b\sqrt{D} \in R : a, b > 0 \text{ and } |\tilde{\alpha}| < 1/b\}.$$

Define $Q \subset \mathbb{Z}^2$ by

$$Q := \{(p, q) : p, q > 0 \text{ and } |\sqrt{D} - p/q| < 1/q^2\}.$$

We claim that these sets are in bijection. Indeed, let $\alpha = a + b\sqrt{D} \in A$. Note that

$$|\sqrt{D} - a/b| = \frac{1}{b}|b\sqrt{D} - a| = \frac{1}{b}|a - b\sqrt{D}| = \frac{1}{b}|\tilde{\alpha}| < 1/b^2.$$

Hence, we can define $f : A \rightarrow Q$ by $f(\alpha) := (a, b)$. Now, let $(p, q) \in Q$. We have that

$$|p - q\sqrt{D}| = |q\sqrt{D} - p| = q|\sqrt{D} - p/q| < 1/q.$$

Hence, we can define $g : Q \rightarrow A$ by $g(p, q) = p + q\sqrt{D}$. Since f and g are mutually inverse, it follows that A and Q are in bijection.

Now, since $\sqrt{D}$ is irrational, we know (by Theorem 2.6) that there are infinitely many rationals p/q satisfying $|\sqrt{D} - p/q| < 1/q^2$. Each such rational corresponds to a pair $(p, q) \in \mathbb{Z}^2$ (with $q > 0$)² satisfying the same condition. Let (p, q) be a pair of this form, and suppose that $p \leq 0$. Then $p/q \leq 0$, and thus

$$|\sqrt{D} - p/q| = \sqrt{D} - p/q \geq \sqrt{D} > 1.$$

But $|\sqrt{D} - p/q| < 1/q^2 \leq 1$, so this is a contradiction. Hence, we must have $p > 0$. Thus, there are infinitely many pairs (p, q) satisfying $|\sqrt{D} - p/q| < 1/q^2$, with $p, q > 0$. That is, the set Q (and hence the set A) is infinite.

We now claim that for each $\alpha \in A$, we have $|N(\alpha)| \leq 1 + 2\sqrt{D}$. Indeed, let $\alpha = a + b\sqrt{D} \in A$. Then

$$\begin{aligned} |N(\alpha)| &= |\alpha\tilde{\alpha}| = |\alpha||\tilde{\alpha}| = |a + b\sqrt{D}||\tilde{\alpha}| = |\tilde{\alpha} + 2b\sqrt{D}||\tilde{\alpha}| < |\tilde{\alpha} + 2b\sqrt{D}|/b \\ &\leq (|\tilde{\alpha}| + 2b\sqrt{D})/b \\ &< 1/b^2 + 2\sqrt{D}. \end{aligned}$$

Since $1/b^2 \leq 1$, the result follows.

For each triple (n, r, s) with $1 \leq n \leq 1 + 2\sqrt{D}$, $0 \leq r, s \leq n - 1$, define $A_{n,r,s}$ by

$$A_{n,r,s} := \{\alpha = a + b\sqrt{D} \in A : |N(\alpha)| = n, a \equiv r \pmod{n} \text{ and } b \equiv s \pmod{n}\}.$$

Note firstly that there are only finitely many such triples, by the boundedness condition. Furthermore, since $N(\alpha) \in \mathbb{Z} \setminus \{0\}$ for each $\alpha \in A$, we have $1 \leq |N(\alpha)| \leq 1 + 2\sqrt{D}$ for each $\alpha \in A$. Thus, for each fixed n , the integers a and b each determine unique residue classes modulo n . Hence $A = \bigcup_{n,r,s} A_{n,r,s}$.

²Recall that this specific characterisation follows from the convention that each such rational is written in reduced form, with strictly positive denominator.

Since the set A is infinite, we can thus conclude that there exists a triple (n, r, s) of the aforementioned form such that $A_{n,r,s}$ is infinite. Let two distinct elements $\alpha, \alpha' \in A_{n,r,s}$ be given, and suppose (without loss of generality) that

$$\alpha = a + b\sqrt{D} > \alpha' = a' + b'\sqrt{D}.$$

We have that $a - a' \equiv 0 \pmod{n}$, $b - b' \equiv 0 \pmod{n}$. Thus, $\gamma := (\alpha - \alpha')/n \in R$. We have $|N(\alpha')| = n \implies n = \pm N(\alpha') = \pm \alpha' \tilde{\alpha}'$. Define $\varepsilon := \alpha/\alpha'$. Then, as $\alpha > \alpha' > 0$, it follows that $\varepsilon > 1$. Furthermore,

$$\varepsilon = \frac{\alpha}{\alpha'} = \frac{\alpha' + n\gamma}{\alpha'} = 1 + \frac{n\gamma}{\alpha'} = 1 + (\pm \tilde{\alpha}')\gamma \in R.$$

Lastly, $|N(\varepsilon)| = |N(\alpha)|/|N(\alpha')| = 1$. Hence, $\varepsilon \in U \setminus \{\pm 1\}$ is a non-trivial unit. $\square$

We can now finish the proof. Note that $\ker L = \{u \in U : |u| = 1\} = \{\pm 1\}$. Now, $L(U)$ is discrete and cyclic. By the proof of Lemma 6.7, if $L(U)$ is non-trivial, then $L(U) = a\mathbb{Z}$, where $a = \inf(L(U) \cap \mathbb{R}^+) > 0$. The fact that $L(U)$ is non-trivial follows by 6.9. Now, since $L(U) = \{ak : k \in \mathbb{Z}\}$, then there exists $\varepsilon \in U$ such that $L(\varepsilon) = a$. Furthermore, this unit must be fundamental, as $L(1) = L(-1) = 0$. Now, let $u \in U$. We then have

$$L(u) = ak = kL(\varepsilon) = L(\varepsilon^k) \implies L(u/\varepsilon^k) = L(u) - L(\varepsilon^k) = 0,$$

for some $k \in \mathbb{Z}$. Hence,

$$|u/\varepsilon^k| = 1 \implies u = \pm \varepsilon^k$$

Since $u \in U$ was taken to be arbitrary, it follows that $U = \{\pm \varepsilon^k : k \in \mathbb{Z}\}$ for some fundamental unit $\varepsilon \notin \{1, -1\}$. This is equivalent to stating that $U \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}$, as the map $\varphi : U \rightarrow \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}$ defined by

$$\varphi(u) = \begin{cases} (0, k) & \text{if } u = \varepsilon^k \text{ for some } k \in \mathbb{Z} \\ (1, k) & \text{if } u = -\varepsilon^k \text{ for some } k \in \mathbb{Z} \end{cases}$$

is an isomorphism. $\square$

We are now able to provide a complete characterisation of the solutions to Pell's equation. Let D be a positive, square-free integer and let $\varepsilon \in U$ be the fundamental unit such that

$$U = \{\pm \varepsilon^k : k \in \mathbb{Z}\}.$$

Suppose that $N(\varepsilon) = 1$. If a pair (x, y) of integers gives a solution to Pell's equation, then $x + y\sqrt{D}$ is a unit. Conversely, let $a + b\sqrt{D} \in U$. By Theorem 6.5, we have $a + b\sqrt{D} = \pm \varepsilon^k$ for some $k \in \mathbb{Z}$. Hence,

$$N(a + b\sqrt{D}) = N(\pm \varepsilon^k) = N(\pm 1)N(\varepsilon)^k = N(\varepsilon)^k = 1.$$

Thus, the pair (a, b) of integers gives a solution to Pell's equation. Hence, solutions to Pell's equation correspond to units of norm 1, and the set

$$\{(x, y) \in \mathbb{Z}^2 : x + y\sqrt{D} \in \{\pm \varepsilon^k : k \in \mathbb{Z}\}\}$$

is precisely the solution set of Pell's equation. Furthermore, the above also implies that the Diophantine equation $x^2 - Dy^2 = -1$ has no integer solutions, since (by our previous

working)

$$x^2 - Dy^2 = -1 \implies N(x + y\sqrt{D}) = -1 \implies N(\varepsilon)^k = -1 \quad \text{for some } k \in \mathbb{Z},$$

a contradiction.

Now, suppose $N(\varepsilon) = -1$. If a pair (x, y) of integers gives a solution to Pell's equation, then $x + y\sqrt{D}$ is again a unit, and $N(x + y\sqrt{D}) = 1$. If a pair (x, y) of integers gives a solution to the Diophantine equation $x^2 - Dy^2 = -1$, then $x + y\sqrt{D}$ is a unit, and $N(x + y\sqrt{D}) = -1$. Conversely, Let $a + b\sqrt{D} \in U$. By an argument similar to that of the $N(\varepsilon) = 1$ case, we then have

$$N(a + b\sqrt{D}) = N(\pm\varepsilon^n) = \begin{cases} 1 & \text{if } n \text{ is even} \\ -1 & \text{if } n \text{ is odd.} \end{cases}$$

Thus, the pair of integers (a, b) gives a solution to Pell's equation if n is even, and a solution to the Diophantine equation $x^2 - Dy^2 = -1$ if n is odd. We then have that

$$\{(x, y) : x + y\sqrt{D} \in \{\pm\varepsilon^{2k} : k \in \mathbb{Z}\}\}$$

is the solution set of Pell's equation and, furthermore,

$$\{(x, y) : x + y\sqrt{D} \in \{\pm\varepsilon^{2k+1} : k \in \mathbb{Z}\}\}$$

is the solution set of the Diophantine equation $x^2 - Dy^2 = -1$.

Before we move on, let us consider some concrete examples. The following table (adapted from [8, pg.127]) displays the structure of $\mathbb{Z}[\sqrt{D}]$ when $D = 5, 7, 11$:

R	U	ε
$\mathbb{Z}[\sqrt{5}]$	$\{\pm(9 + 4\sqrt{5})^k : k \in \mathbb{Z}\}$	$9 + 4\sqrt{5}$
$\mathbb{Z}[\sqrt{7}]$	$\{\pm(8 + 3\sqrt{7})^k : k \in \mathbb{Z}\}$	$8 + 3\sqrt{7}$
$\mathbb{Z}[\sqrt{11}]$	$\{\pm(10 + 3\sqrt{11})^k : k \in \mathbb{Z}\}$	$10 + 3\sqrt{11}$

Table 6.1: Some rings, their unit groups, and fundamental units.

When $D = 5$, for example, we have $N(\varepsilon) = 9^2 - 5 \cdot 4^2 = 1$. By our previous discussion, we can then conclude that the solution set of Pell's equation ($x^2 - 5y^2 = 1$) is precisely

$$\{(x, y) : x + y\sqrt{5} \in \{\pm(9 + 4\sqrt{5})^k : k \in \mathbb{Z}\}\}.$$

Furthermore, we can conclude that the Diophantine equation $x^2 - 5y^2 = -1$ has no integer solutions.

It turns out that the computation of the fundamental unit provides another link to Diophantine approximation. For square-free $\sqrt{D}$, let $[a_0; a_1, a_2, \dots]$ be the sequence of integers in its simple continued fraction expansion. A result of Lagrange [9, Theorem 1.3.8] shows that this sequence is eventually periodic. Furthermore, it is possible to write

$$[a_0; a_1, a_2, \dots] = \left[\lfloor \sqrt{D} \rfloor; \overline{a_1, a_2, \dots, a_{r-1} 2 \lfloor \sqrt{D} \rfloor} \right]$$

for some $r \in \mathbb{N}$ [9, Theorem 1.3.9]. The fundamental unit is then given by

$$\varepsilon := \begin{cases} p_{r-1} + q_{r-1}\sqrt{D} & \text{if } r \text{ is even} \\ p_{r-2} + q_{r-2}\sqrt{D} & \text{if } r \text{ is odd,} \end{cases}$$

where p_i/q_i denotes the i th convergent of the simple continued fraction expansion of $\sqrt{D}$ [8, pg.125]. We omit the proof, since the computations involved are somewhat lengthy. Nevertheless, this connection is testament to the mathematical richness of Diophantine approximation.

6.1 Dirichlet's Unit Theorem

When $D \equiv 2 \pmod{4}$ or $D \equiv 3 \pmod{4}$, Dirichlet's Unit Theorem for Subrings can be attributed to a far more general result of Dirichlet, which describes the structure of the multiplicative group of units $\mathcal{O}_L^\times$ of the ring of algebraic integers $\mathcal{O}_L$ of a number field L . More specifically:

Theorem 6.10 (Dirichlet's Unit Theorem, [10, Theorem 7.4]). *Let L be a number field. Let μ_L be the multiplicative group of roots of unity of L , and $\mathcal{O}_L$ be the ring of algebraic integers of L . Then μ_L is finite and cyclic, and*

$$\mathcal{O}_L^\times \cong \mu_L \times \mathbb{Z}^{r+s-1},$$

where r and $2s$ are the number of real and complex embeddings of L , respectively.

A common anecdote suggests that the idea for the proof of Theorem 6.10 'came to Dirichlet' while he was attending a mass at the Sistine Chapel. While the truth of this story is uncertain, Dirichlet's proof is definitive, and his work underpins much of algebraic number theory.

The structure of the proof does resemble that of the proof of Dirichlet's Unit Theorem for Subrings, but the relevant algebraic structures are inherently more complex. Consider the following statements:

1. There is an additive homomorphism $\ell : L^\times \rightarrow \mathbb{R}^{r+s}$ such that $\ell(\mathcal{O}_L^\times) \subset H$, where H is a subspace of $\mathbb{R}^{r+s}$ such that $\dim H = r + s - 1$.
2. $\ell(\mathcal{O}_L^\times)$ is a lattice in H .
3. $\ker \ell$ is finite and cyclic and, moreover, $\mu_L = \ker \ell$.

Suppose that all of the above statements hold. As $\ell(\mathcal{O}_L^\times)$ is a lattice, then there exist $u_1, \dots, u_{r+s-1} \in \mathcal{O}_L^\times$ such that $\ell(u_1), \dots, \ell(u_{r+s-1})$ form a $\mathbb{Z}$ -basis for $\ell(\mathcal{O}_L^\times)$. Define the map $f : \mu_L \times \mathbb{Z}^{r+s-1} \rightarrow \mathcal{O}_L^\times$ by

$$f(w, a_1, \dots, a_{r+s-1}) := w u_1^{a_1} \cdots u_{r+s-1}^{a_{r+s-1}}.$$

One can then argue that f is an isomorphism, and (3) would allow us to recover the full result.

In order to provide some intuition for the full proof of Dirichlet's Unit Theorem, we can begin by considering the case of number fields of the form $\mathbb{Q}(\sqrt{D})$, where $D > 1$ is square-free. Such fields are often referred to as *real quadratic fields*. The real-quadratic case showcases the fruitful approach of analysing the structure of the field geometrically. The following definitions and results³, for an arbitrary number field L , are necessary for the proof of this special case:

Lemma 6.11 ([11, Lemmas 2.6, 2.8]). *Let $\alpha \in \mathcal{O}_L$. Then $\text{tr}_{L/\mathbb{Q}}(\alpha), N_{L/\mathbb{Q}}(\alpha) \in \mathbb{Z}$. Furthermore, we have $|N_{L/\mathbb{Q}}(\alpha)| = 1$ if and only if $\alpha \in \mathcal{O}_L^\times$.*

³For any statement in this chapter whose proof is not present, a full proof of the statement can be found in the supplementary material.

We will again use the shorthand $N(\alpha)$ for the norm $N_{L/\mathbb{Q}}(\alpha)$ of $\alpha \in \mathcal{O}_L$.

Definition 6.12. Let L be a number field of degree n and let $\sigma_1, \dots, \sigma_n$ denote the embeddings of L into $\mathbb{C}$. Let $\alpha_1, \dots, \alpha_n$ be elements of L . We define the *discriminant* $\text{disc}(\alpha_1, \dots, \alpha_n)$ of the elements $\alpha_1, \dots, \alpha_n$ by

$$\text{disc}(\alpha_1, \dots, \alpha_n) = \det(D)^2,$$

where D is the $n \times n$ matrix defined by $D_{ij} = \sigma_i(\alpha_j)$.

It is worth noting that, as the square is included, the value of the discriminant is invariant under reordering of the elements and/or the embeddings. Indeed, certain reorderings can induce a sign-flip of the determinant, but the square clears this difference.

Definition 6.13. An *integral basis* for $\mathcal{O}_L$ is a tuple $\alpha_1, \dots, \alpha_n$ of elements of L such that

$$\mathcal{O}_L = \bigoplus_{i=1}^n \mathbb{Z}\alpha_i.$$

Definition 6.14. We define $\text{disc}(\mathcal{O}_L) = \text{disc}(\alpha_1, \dots, \alpha_n)$, where $\alpha_1, \dots, \alpha_n$ is any integral basis for $\mathcal{O}_L$.

Lemma 6.15 ([10, Propositions 2.8, 2.10]). *There exists an integral basis for $\mathcal{O}_L$. Furthermore, $\text{disc}(\mathcal{O}_L) \neq 0$.*

Lemma 6.16 ([11, pg.20]). *Let L be a number field of degree $n = r + 2s$. Let $\sigma_1, \dots, \sigma_n$ denote the real embeddings of L , and $\tau_1, \bar{\tau}_1, \dots, \tau_s, \bar{\tau}_s$ denote the s conjugate pairs of complex embeddings. Then the mapping $S : L \rightarrow \mathbb{R}^{r+2s} = \mathbb{R}^r \times \mathbb{C}^s$, defined by*

$$S(\alpha) := (\sigma_1(\alpha), \dots, \sigma_r(\alpha), \Re(\tau_1(\alpha)), \Im(\tau_1(\alpha)), \dots, \Re(\tau_s(\alpha)), \Im(\tau_s(\alpha))).$$

is an additive group homomorphism.

Lemma 6.17 ([11, Lemma 6.10]). *$S(\mathcal{O}_L)$ is a lattice.*

For a lattice $\Lambda \subset \mathbb{R}^n$, we let $A(\Lambda)$ denote the Lebesgue measure of its fundamental parallelepiped. Furthermore, we write vol for the Lebesgue measure in $\mathbb{R}^n$.

Lemma 6.18 ([11, Lemma 6.11]). $A(S(\mathcal{O}_L)) = \frac{1}{2^s} \sqrt{|\text{disc}(\mathcal{O}_L)|}$.

Corollary 6.19 ([11, Lemma 7.6]). *Let $B > 0$ be a real number, and let*

$$X_B := \{\alpha \in \mathcal{O}_L : \text{for all embeddings } \sigma : L \rightarrow \mathbb{C}, |\sigma(\alpha)| \leq B\}.$$

Then X_B is finite.

Proof. Note that

$$S(X_B) = S(\mathcal{O}_L) \cap [-B, B]^r \times \{z \in \mathbb{C}^s : |z_j| \leq B \text{ for each } j \in \{1, \dots, s\}\},$$

and the set on the right-hand side is compact. Thus, as $S(\mathcal{O}_L)$ is a lattice, it follows that $S(X_B)$ is finite. Since S is injective, it follows that X_B is finite. $\square$

Theorem 6.20 (Minkowski's Theorem, [10, Theorem 4.4]). *Let Λ be a lattice in $\mathbb{R}^n$, and let $E \subset \mathbb{R}^n$ be a convex, centrally symmetric subset. Then if $\text{vol}(E) > 2^n A(\Lambda)$, E contains a non-zero point of Λ . If E is compact, the conclusion holds under the weaker assumption that $\text{vol}(E) \geq 2^n A(\Lambda)$.*

Theorem 6.21 (Dirichlet's Unit Theorem for Real Quadratic Fields). *Let $D > 1$ be square-free. Write L for the field $\mathbb{Q}(\sqrt{D})$. Then there exists some $\varepsilon \in \mathcal{O}_L^\times$ such that*

$$\mathcal{O}_L^\times = \{\pm \varepsilon^k : k \in \mathbb{Z}\}.$$

Proof (adapted from [12]). Note firstly that there are two real embeddings of L . Namely, they are the inclusion $\iota : \mathbb{Q}(\sqrt{D}) \hookrightarrow \mathbb{C}$, and the conjugation map $c : \mathbb{Q}(\sqrt{D}) \rightarrow \mathbb{C}; \alpha \mapsto \tilde{\alpha}$. Thus, $r = 2$ and $s = 0$. Let $t_1 > 0$ and consider the set

$$s_{t_1} = \left\{ (y_1, y_2) \in \mathbb{R}^2 : |y_1| \leq t_1, |y_2| \leq \frac{|\text{disc}(\mathcal{O}_L)|^{1/2}}{t_1} \right\}.$$

Note that we have $\text{vol}(s_{t_1}) = 4|\text{disc}(\mathcal{O}_L)|^{1/2} = 2^2 A(S(\mathcal{O}_L))$. Furthermore, s_{t_1} is convex and centrally symmetric. Since it is also closed and bounded, it is compact by the Heine-Borel Theorem. Hence, by Minkowski's Theorem, we have that there exists some non-zero $\alpha \in \mathcal{O}_L$ such that $S(\alpha) \in s_{t_1}$. Since α is non-zero, the norm is integer-valued and $S(\alpha) = (\alpha, \tilde{\alpha})$, it follows that $1 \leq |N(\alpha)| \leq |\text{disc}(\mathcal{O}_L)|^{1/2}$ for any such point. Now, s_{t_1} is compact and must thus intersect $S(\mathcal{O}_L)$ in a finite set. Furthermore, none of the non-zero points in $S(\mathcal{O}_L)$ lie on either of the coordinate axes, since $\sqrt{D}$ is irrational. Hence, there exists some non-zero t_2 such that $|y_1| > t_2$ for all $(y_1, y_2) \in s_{t_1} \cap S(\mathcal{O}_L)$. Now, we again have that there must exist some non-zero $\alpha \in s_{t_2} \cap \mathcal{O}_L$. Moreover, any such point must satisfy $1 \leq |N(\alpha)| \leq |\text{disc}(\mathcal{O}_L)|^{1/2}$, and we have $s_{t_2} \cap (s_{t_1} \cap S(\mathcal{O}_L)) = \emptyset$. We can continue this process inductively. That is, given $t_1 > \dots > t_n$, we can choose t_{n+1} such that

$$0 < t_{n+1} < \min \left\{ |y_1| : (y_1, y_2) \in \bigcup_{i=1}^n s_{t_i} \cap S(\mathcal{O}_L) \right\},$$

and we have that there must exist some non-zero $\alpha \in s_{t_{n+1}} \cap \mathcal{O}_L$ such that $1 \leq |N(\alpha)| \leq |\text{disc}(\mathcal{O}_L)|^{1/2}$ and $\alpha \notin s_{t_i}$ for any $i < n + 1$. This gives us infinitely many $\alpha \in \mathcal{O}_L$ such that $1 \leq |N(\alpha)| \leq |\text{disc}(\mathcal{O}_L)|^{1/2}$. As $N(\alpha) \in \mathbb{Z}$, the pigeonhole principle implies that there must exist some $m \in \mathbb{Z}$ such that $N(\alpha) = m$ for infinitely many $\alpha \in \mathcal{O}_L$. Now, the map $\mathcal{O}_L \rightarrow (\mathbb{Z}/m\mathbb{Z})^2; a + b\sqrt{D} \mapsto (a \bmod m, b \bmod m)$ defines a surjective group homomorphism with kernel $m\mathcal{O}_L$. Thus,

$$\mathcal{O}_L/m\mathcal{O}_L \cong (\mathbb{Z}/m\mathbb{Z})^2,$$

another finite set. Hence, we see that each $\alpha \in \mathcal{O}_L$ must belong to exactly one (of the finitely many) cosets of $m\mathcal{O}_L$ in $\mathcal{O}_L$. By the pigeonhole principle there must exist a non-zero $\beta \in \mathcal{O}_L$ with $N(\beta) = m$ and infinitely many non-zero $\alpha \in \mathcal{O}_L$ with $N(\alpha) = m$ and $\alpha m\mathcal{O}_L = \beta m\mathcal{O}_L$. For each such α , there exists $\gamma \in \mathcal{O}_L$ such that

$$\alpha = \beta + m\gamma \implies u := \frac{\alpha}{\beta} = 1 + \frac{m}{\beta}\gamma = 1 + \frac{N(\beta)}{\beta}\gamma = 1 + \tilde{\beta}\gamma \in \mathcal{O}_L,$$

and $|N(\alpha/\beta)| = |N(\alpha)|/|N(\beta)| = 1$. Since $\alpha/\beta = \alpha'/\beta \iff \alpha = \alpha'$, then the above implies that we have constructed infinitely many units $u \in \mathcal{O}_L^\times$. Hence, there must exist some $\varepsilon_0 \in \mathcal{O}_L^\times$ such that $\varepsilon_0 \notin \{\pm 1\}$. Recall that $S(\varepsilon_0) = (\iota(\varepsilon_0), c(\varepsilon_0))$. Since ε_0^{-1} is also a

unit, we have that

$$|\iota(\varepsilon_0^{-1})| = \frac{1}{|\iota(\varepsilon_0)|} \implies |\iota(\varepsilon_1)| > 1 \text{ for some } \varepsilon_1 \in \{\varepsilon_0, \varepsilon_0^{-1}\} \subset \mathcal{O}_L^\times,$$

as $\varepsilon_0 \notin \{\pm 1\}$. Define

$$X_1 := \{\alpha \in \mathcal{O}_L^\times : 1 \leq |\iota(\alpha)| \leq |\varepsilon_1|\} = \{\alpha \in \mathcal{O}_L : |N(\alpha)| = 1, 1 \leq |\iota(\alpha)| \leq |\varepsilon_1|\}.$$

By Corollary 6.19, this set is finite. Hence, there exists $\varepsilon_2 \in X_1 \setminus \{\pm 1\}$ such that $|\iota(\varepsilon_2)| \leq |\alpha|$ for all $\alpha \in X_1 \setminus \{\pm 1\}$. Now, since $|\iota(-\varepsilon_2)| = |\iota(\varepsilon_2)|$, then we can choose $\varepsilon \in \{\pm \varepsilon_2\}$ such that $\iota(\varepsilon) > 0$. Now, let $\alpha \in \mathcal{O}_L^\times$, and assume that $\iota(\alpha) > 0$. Then, for some $k \in \mathbb{N}$ and $0 \leq \gamma < 1$, we can write ⁴

$$\begin{aligned} \frac{\log(\iota(\alpha))}{\log(\iota(\varepsilon))} &= k + \gamma \implies \log(\iota(\alpha)) = (k + \gamma) \log(\iota(\varepsilon)) = \log(\iota(\varepsilon)^{k+\gamma}) \\ &\implies \iota(\alpha) = \iota(\varepsilon)^{k+\gamma} = \iota(\varepsilon^k) \iota(\varepsilon)^\gamma \\ &\implies \iota(\alpha \varepsilon^{-k}) = \iota(\varepsilon)^\gamma. \end{aligned}$$

Note that $\alpha \varepsilon^{-k} \in \mathcal{O}_L^\times$. Furthermore, $|\iota(\varepsilon)^\gamma| = |\iota(\varepsilon)|^\gamma < |\iota(\varepsilon)|$, as $|\iota(\varepsilon)| > 1$. Note then that, if $\gamma \neq 0$, this contradicts minimality of ε . Hence, we must have

$$\iota(\alpha \varepsilon^{-k}) = 1 \implies \alpha \varepsilon^{-k} = 1 \implies \alpha = \varepsilon^k.$$

Now, let $\beta \in \mathcal{O}_L^\times$ be given. By the above, if $\iota(\beta) > 0$, then $\beta = \varepsilon^k$ for some $k \in \mathbb{N}$. Otherwise, $\iota(-\beta) > 0$ and $\beta = -\varepsilon^k$ for some $k \in \mathbb{N}$. Hence, we have that

$$\mathcal{O}_L^\times = \{\pm \varepsilon^k : k \in \mathbb{Z}\},$$

as required. □

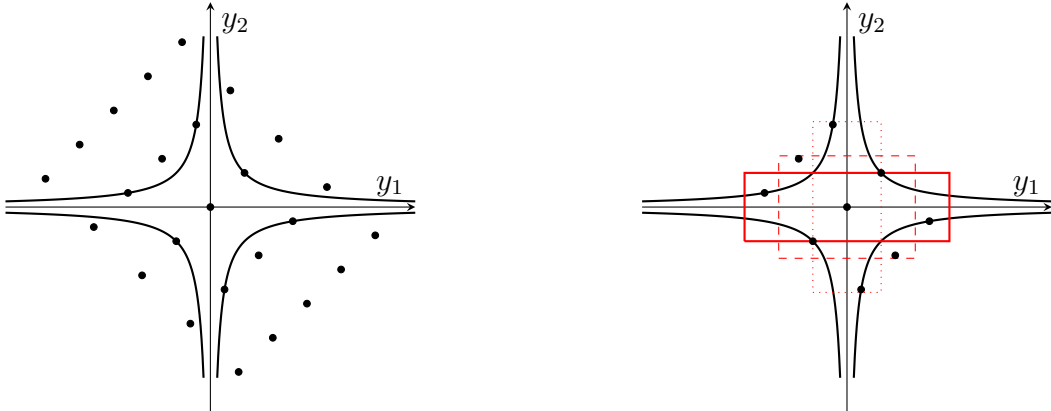


Figure 6.1: Left: The embedding of $\mathcal{O}_L$ as a lattice in $\mathbb{R}^2$ with the curve $|y_1 y_2| = 1$, where $L = \mathbb{Q}(\sqrt{2})$. Right: The same curve together with the sets s_{t_1} (solid border), s_{t_2} (dashed border), s_{t_3} (dotted border), each containing a new lattice point. Here $t_1 = 3$, $t_2 = 2$, $t_3 = 1.2$ (adapted from [13, pg. 5])

⁴It is here that the use of ι is important. Strictly speaking, fractional powers of ε do not make sense in the ring $\mathcal{O}_L$.

Consider the following table, which displays some real quadratic fields and the explicit computations of their unit groups. Notice that the ring of algebraic integers of $\mathbb{Q}(\sqrt{5})$ is not $\mathbb{Z}[\sqrt{5}]$. This is actually the result of a much more general fact; if $D > 1$ is square-free, then $\mathcal{O}_{\mathbb{Q}(\sqrt{D})} = \mathbb{Z}[\sqrt{D}]$ if and only if $D \equiv 2 \pmod{4}$ or $D \equiv 3 \pmod{4}$. Otherwise, we have $\mathcal{O}_{\mathbb{Q}(\sqrt{D})} = \mathbb{Z}[\frac{1+\sqrt{D}}{2}]$. Thus, the more ‘general’ result comes with a caveat:

L	$\mathcal{O}_L$	$\mathcal{O}_L^\times$
$\mathbb{Q}(\sqrt{2})$	$\mathbb{Z}[\sqrt{2}]$	$\{\pm(1 + \sqrt{2})^k : k \in \mathbb{Z}\}$
$\mathbb{Q}(\sqrt{3})$	$\mathbb{Z}[\sqrt{3}]$	$\{\pm(2 + \sqrt{3})^k : k \in \mathbb{Z}\}$
$\mathbb{Q}(\sqrt{5})$	$\mathbb{Z}[\frac{1+\sqrt{5}}{2}]$	$\{\pm(\frac{1+\sqrt{5}}{2})^k : k \in \mathbb{Z}\}$

Table 6.2: Some real quadratic fields, their rings of integers, and unit groups. (adapted from [13, pg. 2])

6.2 Proof of Dirichlet’s Unit Theorem

We proceed with the proof of Dirichlet’s Unit Theorem, in accordance with the structure discussed in the previous chapter.

Lemma 6.22 ([10, Proposition 2.6]). *Let L/K be an extension of number fields of degree $[L : K] = n$, and let $\sigma_0 : K \rightarrow \mathbb{C}$ be a complex embedding. Let $\sigma_1, \dots, \sigma_n$ be the distinct complex embeddings such that $\sigma_i|_K = \sigma_0$ for each $i \in \{1, \dots, n\}$. Then, for each $\alpha \in L$, we have*

$$\sigma_0(\text{tr}_{L/K}(\alpha)) = \sum_{i=1}^n \sigma_i(\alpha), \quad \sigma_0(N_{L/K}(\alpha)) = \prod_{i=1}^n \sigma_i(\alpha).$$

Lemma 6.23 ([11, pg. 22]). *There is an additive homomorphism $\ell : L^\times \rightarrow \mathbb{R}^{r+s}$ such that $\ell(\mathcal{O}_L^\times) \subset H$, where H is a subspace of $\mathbb{R}^{r+s}$ such that $\dim H = r + s - 1$.*

Proof. Let L be a number field. Let $\sigma_1, \dots, \sigma_r, \tau_1, \overline{\tau}_1, \dots, \tau_s, \overline{\tau}_s$ denote the complex embeddings of L . Consider the map $\ell : L^\times \rightarrow \mathbb{R}^{r+s}$, defined by

$$\ell(\alpha) = (\log|\sigma_1(\alpha)|, \dots, \log|\sigma_r(\alpha)|, 2\log|\tau_1(\alpha)|, \dots, 2\log|\tau_s(\alpha)|).$$

The multiplicativity of the embeddings, together with the identity $\log(xy) = \log x + \log y$, implies that ℓ defines a group homomorphism. Now, we claim that

$$\ell(\mathcal{O}_L^\times) \subseteq H = \left\{ (x_1, \dots, x_{r+s}) : \sum_{i=1}^{r+s} x_i = 0 \right\}$$

Indeed, let $\alpha \in \mathcal{O}_L^\times$. By Lemma 6.22, taking $\sigma_0 : \mathbb{Q} \hookrightarrow \mathbb{C}$ to be the inclusion, we have

$$\begin{aligned} N_{L/\mathbb{Q}}(\alpha) &= \prod_{i=1}^r \sigma_i(\alpha) \prod_{i=1}^s \tau_i(\alpha) \overline{\tau}_i(\alpha) \\ &= \prod_{i=1}^r \sigma_i(\alpha) \prod_{i=1}^s |\tau_i(\alpha)|^2 = 1. \end{aligned}$$

Taking the log of the absolute values, we have

$$\begin{aligned} \log|N_{L/\mathbb{Q}}(\alpha)| &= \sum_{i=1}^r \log|\sigma_i(\alpha)| + 2 \sum_{i=1}^s \log|\tau_i(\alpha)| = 0 \\ \implies \ell(\alpha) &\in H. \end{aligned}$$

Thus, $\ell(\mathcal{O}_L^\times) \subseteq H$.

Now, note that H is a subspace of $\mathbb{R}^{r+s}$ (the summation property of the elements of H is invariant under scalar multiplication, and H is clearly closed under addition). Note also that $H = \ker \Sigma$, where $\Sigma : \mathbb{R}^{r+s} \rightarrow \mathbb{R}$ is defined by

$$\Sigma(x_1, \dots, x_{r+s}) = \sum_{i=1}^{r+s} x_i.$$

Furthermore, it is clear that Σ is a linear mapping. Note also that Σ is surjective ($\Sigma(t, 0, \dots, 0) = t$ for any $t \in \mathbb{R}$), and thus $\dim \operatorname{Im} \Sigma = \dim \mathbb{R} = 1$. Hence, by the rank-nullity theorem, we have that

$$\dim H = \dim \ker \Sigma = \dim(\mathbb{R}^{r+s}) - \dim(\operatorname{Im} \Sigma) = r + s - 1.$$

Thus, H is a subspace of $\mathbb{R}^{r+s}$ of dimension $r + s - 1$. $\square$

Lemma 6.24 ([14, pg. 102]). *Let $N \geq 1$ be an integer and let $A \in M_N(\mathbb{R})$ be a matrix satisfying the following conditions:*

1. *For each $j \in \{1, \dots, N\}$, $\sum_{i=1}^N A_{ij} = 0$.*
2. *For all $i, j \in \{1, \dots, N\}$, we have $A_{ij} > 0$ if $i = j$ and $A_{ij} < 0$ if $i \neq j$.*

Then A has rank $N - 1$.

Lemma 6.25 ([14, pg. 102]). *Let $1 \leq k \leq r + s$ be an integer. Then there exists an element $\varepsilon \in \mathcal{O}_L^\times$ such that, writing $\ell(\varepsilon) = (e_1, \dots, e_{r+s})$, we have $e_i > 0$ if $i \neq k$ and $e_k < 0$.*

Lemma 6.26 (Sandwich Lemma, [11, Lemma 3.7]).

1. *Let $H \subset G$ be abelian groups such that $G \cong \mathbb{Z}^n$ for some $n \geq 1$. Then $H \cong \mathbb{Z}^m$ for some $m \leq n$.*
2. *Let $K \subset H \subset G$ be abelian groups such that $K \cong \mathbb{Z}^n$ and $G \cong \mathbb{Z}^n$ for some $n \geq 1$. Then $H \cong \mathbb{Z}^n$.*
3. *Let $H \subset G$ be abelian groups such that $H \cong G \cong \mathbb{Z}^n$. Then G/H is finite.*

Lemma 6.27 ([10, Theorem 7.3]). *$\ell(\mathcal{O}_L^\times)$ is a lattice in H .*

Proof. We first show that $\ell(\mathcal{O}_L^\times)$ spans H . By Lemma 6.25, there exist $v_1, \dots, v_{r+s} \in \ell(\mathcal{O}_L^\times)$ such that the i th entry of v_j is strictly positive if $i \neq j$, and negative otherwise, for each $j \in \{1, \dots, r + s\}$.

Let $A \in M_{r+s}(\mathbb{R})$ be the matrix with column j given by v_j . By Lemma 6.24, the rank of A is $r + s - 1$. Earlier, we computed that $\dim H = r + s - 1$, and thus $\ell(\mathcal{O}_L^\times)$ spans H .

Recall that any spanning set of a finite-dimensional vector space contains a basis. Thus, we may choose vectors $v_1, \dots, v_{r+s-1} \in \ell(\mathcal{O}_L^\times)$ that form a basis of H as a vector space over $\mathbb{R}$. Define

$$\Lambda = \bigoplus_{i=1}^{r+s-1} \mathbb{Z}v_i.$$

Then $v_1, \dots, v_{r+s-1}$ span Λ , and $\Lambda \subset \ell(\mathcal{O}_L^\times)$. Let $P \subset H$ be defined by

$$P = \left\{ \sum_{i=1}^{r+s-1} t_i v_i : t_i \in [0, 1] \text{ for each } i \in \{1, \dots, r+s-1\} \right\}.$$

Now, suppose that $\ell(\alpha) \in P$ for some $\alpha \in \mathcal{O}_L^\times$. Then, by the definition of ℓ , $|\sigma(\alpha)|$ is bounded for any embedding $\sigma : L \rightarrow \mathbb{C}$, and furthermore this bound is independent of the embedding (as P is bounded). It follows, by Corollary 6.19, that $P \cap \ell(\mathcal{O}_L^\times)$ is finite.

Let $x \in \ell(\mathcal{O}_L^\times)$. Write

$$x = \sum_{i=1}^{r+s-1} a_i v_i$$

for some $a_1, \dots, a_{r+s-1} \in \mathbb{R}$. Then, we have that

$$x = \sum_{i=1}^{r+s-1} ([a_i] + t_i) v_i = \sum_{i=1}^{r+s-1} [a_i] v_i + \sum_{i=1}^{r+s-1} t_i v_i.$$

for some $t_1, \dots, t_{r+s-1} \in [0, 1]$. Since $\lambda = \sum_{i=1}^{r+s-1} [a_i] v_i \in \Lambda$, $p = \sum_{i=1}^{r+s-1} t_i v_i \in P$, it follows that we can write $x = \lambda + p$ for some $\lambda \in \Lambda$ and $p \in P$, for each $x \in \ell(\mathcal{O}_L^\times)$.

Note then that, for each $x \in \ell(\mathcal{O}_L^\times)$,

$$p = (\lambda - x) \in \ell(\mathcal{O}_L^\times) \cap P,$$

as $\Lambda \subset \ell(\mathcal{O}_L^\times)$, and $\ell(\mathcal{O}_L^\times)$ is a group. Hence, for each $x \in \ell(\mathcal{O}_L^\times)$, we can write $x = \lambda + p$ for some $p \in \ell(\mathcal{O}_L^\times) \cap P$. As $\ell(\mathcal{O}_L^\times) \cap P$ is finite, this means that the number of cosets of $\Lambda \subset \ell(\mathcal{O}_L^\times)$ is finite. Note that we can consider the quotient group $\ell(\mathcal{O}_L^\times)/\Lambda$, as $\ell(\mathcal{O}_L^\times)$ is abelian, so all of its subgroups are normal. Hence, we can write $[\ell(\mathcal{O}_L^\times) : \Lambda] = N$ for some $N \in \mathbb{N}$. By Lagrange's Theorem, we then have that $N(x + \Lambda) = 0_{\ell(\mathcal{O}_L^\times)}$, for any $x \in \ell(\mathcal{O}_L^\times)$. This is equivalent to stating that $N\ell(\mathcal{O}_L^\times) \subset \Lambda$. Hence, we have that

$$\Lambda \subset \ell(\mathcal{O}_L^\times) \subset \frac{1}{N}\Lambda.$$

Hence, by the Sandwich Lemma, $\ell(\mathcal{O}_L^\times) \cong \mathbb{Z}^{r+s-1}$. That is, $\ell(\mathcal{O}_L^\times)$ is a lattice in H . $\square$

Note that $\ker \ell \subseteq X_1$, where X_1 is defined as in Corollary 6.19. Thus, $\ker \ell$ is finite. Hence, each element of $\ker \ell$ is of finite order. Thus, $\ker \ell \subseteq \mu_L$. Conversely, if $\alpha \in \mu_L$, then $\alpha^N = 1$ for some $N \in \mathbb{N}$, and (as ℓ is a group homomorphism, the identity element $1 \in L^\times$ is mapped to 0)

$$\ell(\alpha^N) = N\ell(\alpha) = 0 \implies \ell(\alpha) = 0.$$

It follows that $\mu_L \subseteq \ker \ell$ and hence that $\mu_L = \ker \ell$. Thus, μ_L is finite and hence cyclic, as any finite subgroup of the group of roots of unity in $\mathbb{C}$ is cyclic. Now, let $u_1, \dots, u_{r+s-1} \in \mathcal{O}_L^\times$ be elements whose image under ℓ forms a $\mathbb{Z}$ -basis of $\ell(\mathcal{O}_L^\times)$. Let $f : \mu_L \times \mathbb{Z}^{r+s-1} \rightarrow \mathcal{O}_L^\times$ be the mapping defined by

$$f(w, a_1, \dots, a_{r+s-1}) = w u_1^{a_1} \cdots u_{r+s-1}^{a_{r+s-1}}.$$

Firstly,

$$f((w_1, a_1, \dots, a_{r+s-1}) \cdot (w_2, b_1, \dots, b_{r+s-1})) = (w_1 w_2 u_1^{a_1+b_1} \cdots u_{r+s-1}^{a_{r+s-1}+b_{r+s-1}})$$

$$\begin{aligned}
 &= (w_1 u_1^{a_1} \cdots u_{r+s-1}^{a_{r+s-1}})(w_2 u_1^{b_1} \cdots u_{r+s-1}^{b_{r+s-1}}) \\
 &= f(w_1, a_1, \dots, a_{r+s-1})f(w_2, b_1, \dots, b_{r+s-1}),
 \end{aligned}$$

so f is a homomorphism. Note that the second step above uses the fact that $\mathcal{O}_L^\times$ is abelian. Furthermore, suppose $(w, a_1, \dots, a_{r+s-1}) \in \ker f$. Then, we have that

$$0 = \ell(1) = \ell(w u_1^{a_1} \cdots u_{r+s-1}^{a_{r+s-1}}) = \ell(w) + \sum_{i=1}^{r+s-1} \ell(u_i^{a_i}) = \sum_{i=1}^{r+s-1} a_i \ell(u_i).$$

Since $u_1, \dots, u_{r+s-1} \in \mathcal{O}_L^\times$ are elements whose image under ℓ forms a $\mathbb{Z}$ -basis of $\ell(\mathcal{O}_L^\times)$, we must have $a_i = 0$ for all $i \in \{1, \dots, r+s-1\}$. Hence, $\ker f = (1, 0, \dots, 0)$. Thus, only element of $\ker f$ is the identity, and f is thus injective. Finally, let $\alpha \in \mathcal{O}_L^\times$ be given. Then $\ell(\alpha) = \sum_{i=1}^{r+s-1} b_i \ell(u_i)$, for some $b_1, \dots, b_{r+s-1} \in \mathbb{Z}$. Let $w = \alpha \prod_{i=1}^{r+s-1} u_i^{-b_i}$. Then

$$\ell(w) = \ell(\alpha) - \sum_{i=1}^{r+s-1} b_i \ell(u_i) = 0,$$

and $w \in \ker \ell = \mu_L$. Note then that

$$f(w, b_1, \dots, b_{r+s-1}) = w u_1^{b_1} \cdots u_{r+s-1}^{b_{r+s-1}} = \alpha,$$

and f is thus surjective. Hence, f is an isomorphism, and the proof of Dirichlet's Unit Theorem is complete. $\square$

We now have a complete classification of the structure of the unit group $\mathcal{O}_L^\times$ of a finite field extension L over $\mathbb{Q}$, which leads to some more exotic applications:

L	$\mathcal{O}_L$	$\mathcal{O}_L^\times$
$\mathbb{Q}(\zeta_5)$	$\mathbb{Z}[\zeta_5]$	$\mu_{10} \times \mathbb{Z}$
$\mathbb{Q}(i)$	$\mathbb{Z}[i]$	$\{\pm 1, \pm i\}$
$\mathbb{Q}(\sqrt[3]{2})$	$\mathbb{Z}[\sqrt[3]{2}]$	$\{\pm 1\} \times \mathbb{Z}$

Table 6.3: Some more number fields, their rings of integers, and unit groups. Here ζ_5 denotes the primitive 5th root of unity, and μ_{10} denotes the cyclic group $\{e^{2\pi i k/10} : k = 0, \dots, 9\}$ (adapted from [13, pg. 2])

Although the connection between Dirichlet's Unit Theorem and Diophantine approximation is less obvious, its inception was certainly motivated by this field of study. Pell's equation has been a catalyst of mathematical progression since antiquity, when mathematicians in India and Greece sought to approximate $\sqrt{2}$. This motivated much of the study of real quadratic fields. As humanity progressed and mathematics grew, so too did the theory of field extensions over $\mathbb{Q}$. This led to Dirichlet's magnificent result, which metaphorically swept the board in the effort to understand the structure of the unit groups of number fields. This is further testament to the mathematical influence of the study of rational approximation.

Chapter 7

Conclusion

Diophantine Approximation has encompassed the minds of mathematicians since the Common Era. Fascinatingly, it remains prevalent in contemporary mathematical research. James Maynard was awarded the Fields Medal in 2022 for work closely connected to this area. The work of Liouville, Thue, and Roth underpins much of the modern work on rational approximation. Together, they pioneered the transformation of the initial question of rational approximability into a deep and far-reaching theory.

In this report, we have mapped out the development of the study of rational approximation. Specifically, we have showcased a chain of successive results that grew increasingly influential as the nuance of rational approximation emerged. Furthermore, these results employed a wide range of techniques, and allowed us to gain appreciation for the origin and subsequent development of the polynomial method.

As the bounds on approximation fastened, the methods of proof became increasingly sophisticated. While the proof of Roth's Theorem is far beyond the scope of this report, we have been able to illustrate its importance and understand its implications. In particular, we introduced applications of the work of Thue and Liouville that allude to the weight of Roth's result and the brilliance of its definitive nature.

Despite the fact that this report has mainly focused on classical results in Diophantine approximation, modern work continues to build on its foundations. The enduring vitality of the field is testament to the depth of the questions that sparked its inception and the minds of those who first sought to answer them.

Theorem 7.1 (Roth, 1955 [1, Theorem 1.2]). *Any algebraic number β of degree $d \geq 2$ is not approximable to any order greater than 2.*

Bibliography

- [1] Michael Nakamaye. Diophantine Approximation on Algebraic Varieties. 11(2):439–502.
- [2] Larry Guth. Lecture notes: Introduction to Diophantine Equations.
- [3] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, 4 edition.
- [4] Jim Wright. Project tutorials on Diophantine Approximation.
- [5] Adam Grabowski and Artur Korniłowicz. Introduction to Liouville Numbers. 25(1):39–48.
- [6] M. Ram Murty and Purusottam Rath. *Transcendental Numbers*. Springer.
- [7] Tom Leinster. Galois Theory.
- [8] Titu Andreescu, Dorin Andrica, and Ion Cucurezeanu. *An Introduction to Diophantine Equations*. Birkhäuser Boston.
- [9] Jörn Steuding, editor. *Diophantine Analysis*. Trends in Mathematics. Birkhäuser.
- [10] Jürgen Neukirch. *Algebraic Number Theory*, volume 322 of *Grundlehren der mathematischen Wissenschaften*. Springer.
- [11] Jack A. Thorne. Number fields. Lecture notes for the Part II Number Fields course, University of Cambridge.
- [12] Dexter Chua. Part II Number Fields – Dirichlet’s Unit Theorem. Cambridge lecture notes.
- [13] Keith Conrad. Dirichlet’s Unit Theorem. Expository note.
- [14] Daniel A. Marcus. *Number Fields*. Universitext. Springer.