Geometric Number Theory

Lenny Fukshansky
Minkowki’s creation of the geometry of numbers was likened to the story of Saul, who set out to look for his father’s asses and discovered a Kingdom.

J. V. Armitage
## Contents

Chapter 1. Geometry of Numbers 1
1.1. Introduction 1
1.2. Lattices 2
1.3. Theorems of Blichfeldt and Minkowski 10
1.4. Successive minima 13
1.5. Inhomogeneous minimum 18
1.6. Problems 21

Chapter 2. Discrete Optimization Problems 23
2.1. Sphere packing, covering and kissing number problems 23
2.2. Lattice packings in dimension 2 29
2.3. Algorithmic problems on lattices 34
2.4. Problems 38

Chapter 3. Quadratic forms 39
3.1. Introduction to quadratic forms 39
3.2. Minkowski’s reduction 46
3.3. Sums of squares 49
3.4. Problems 53

Chapter 4. Diophantine Approximation 55
4.1. Real and rational numbers 55
4.2. Algebraic and transcendental numbers 57
4.3. Dirichlet’s Theorem 61
4.4. Liouville’s theorem and construction of a transcendental number 65
4.5. Roth’s theorem 67
4.6. Continued fractions 70
4.7. Kronecker’s theorem 76
4.8. Problems 79

Chapter 5. Algebraic Number Theory 82
5.1. Some field theory 82
5.2. Number fields and rings of integers 88
5.3. Noetherian rings and factorization 97
5.4. Norm, trace, discriminant 101
5.5. Fractional ideals 105
5.6. Further properties of ideals 109
5.7. Minkowski embedding 113
5.8. The class group 116
5.9. Dirichlet’s unit theorem 119
# CONTENTS

5.10. Problems \hspace{1cm} 120

## Appendices

Appendix A. Some properties of abelian groups \hspace{1cm} 124

Appendix B. Maximum Modulus Principle and Fundamental Theorem of Algebra \hspace{1cm} 127

Appendix. Bibliography \hspace{1cm} 129
CHAPTER 1

Geometry of Numbers

1.1. Introduction

The foundations of the Geometry of Numbers were laid down by Hermann Minkowski in his monograph “Geometrie der Zahlen”, which was published in 1910, a year after his death. This subject is concerned with the interplay of compact convex 0-symmetric sets and lattices in Euclidean spaces. A set \( K \subset \mathbb{R}^n \) is compact if it is closed and bounded, and it is convex if for any pair of points \( x, y \in K \) the line segment connecting them is entirely contained in \( K \), i.e. for any \( 0 \leq t \leq 1 \), \( tx + (1 - t)y \in K \). Further, \( K \) is called 0-symmetric if for any \( x \in K \), \(-x \in K\).

Given such a set \( K \) in \( \mathbb{R}^n \), one can ask for an easy criterion to determine if \( K \) contains any nonzero points with integer coordinates. While for an arbitrary set \( K \) such a criterion can be rather difficult, in case of \( K \) as above a criterion purely in terms of its volume is provided by Minkowski’s fundamental theorem.

It is not difficult to see that \( K \) must in fact be convex and 0-symmetric for a criterion like this purely in terms of the volume of \( K \) to be possible. Indeed, the rectangle

\[
R = \{(x, y) \in \mathbb{R}^2 : 1/3 \leq x \leq 2/3, -t \leq y \leq t\}
\]

is convex for every \( t \), but not 0-symmetric, and its area is \( 2t/3 \), which can be arbitrarily large depending on \( t \) while it still contains no integer points at all. On the other hand, the set \( R^+ \cup -R^+ \) where

\[
R^+ = \{(x, y) \in R : y \geq 0\}
\]

and \(-R^+ = \{(-x, -y) : (x, y) \in R^+\}\) is 0-symmetric, but not convex, and again can have arbitrarily large area while containing no integer points.

Minkowski’s theory applies not only to the integer lattice, but also to more general lattices. Our goal in this chapter is to introduce Minkowski’s powerful theory, starting with the basic notions of lattices.
1.2. Lattices

We start with an algebraic definition of lattices. Let $\mathbf{a}_1, \ldots, \mathbf{a}_r$ be a collection of linearly independent vectors in $\mathbb{R}^n$.

**Definition 1.2.1.** A lattice $\Lambda$ of rank $r$, $1 \leq r \leq n$, spanned by $\mathbf{a}_1, \ldots, \mathbf{a}_r$ in $\mathbb{R}^n$ is the set of all possible linear combinations of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_r$ with integer coefficients. In other words, 

$$\Lambda = \text{span}_\mathbb{Z} \{ \mathbf{a}_1, \ldots, \mathbf{a}_r \} := \left\{ \sum_{i=1}^{r} n_i \mathbf{a}_i : n_i \in \mathbb{Z} \text{ for all } 1 \leq i \leq r \right\}.$$ 

The set $\mathbf{a}_1, \ldots, \mathbf{a}_r$ is called a basis for $\Lambda$. There are usually infinitely many different bases for a given lattice.

Notice that in general a lattice in $\mathbb{R}^n$ can have any rank $1 \leq r \leq n$. We will however talk specifically about lattices of rank $n$, that is of full rank. The most obvious example of a lattice is the set of all points with integer coordinates in $\mathbb{R}^n$: $\mathbb{Z}^n = \{ \mathbf{x} = (x_1, \ldots, x_n) : x_i \in \mathbb{Z} \text{ for all } 1 \leq i \leq n \}$.

Notice that the set of standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, where 

$$\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0),$$

with 1 in $i$-th position is a basis for $\mathbb{Z}^n$. Another basis is the set of all vectors

$$\mathbf{e}_i + \mathbf{e}_{i+1}, \ 1 \leq i \leq n - 1.$$

If $\Lambda$ is a lattice of rank $r$ in $\mathbb{R}^n$ with a basis $\mathbf{a}_1, \ldots, \mathbf{a}_r$ and $\mathbf{y} \in \Lambda$, then there exist $m_1, \ldots, m_r \in \mathbb{Z}$ such that

$$\mathbf{y} = \sum_{i=1}^{r} m_i \mathbf{a}_i = A \mathbf{m},$$

where

$$\mathbf{m} = \begin{pmatrix} m_1 \\ \vdots \\ m_r \end{pmatrix} \in \mathbb{Z}^r,$$

and $A$ is an $n \times r$ basis matrix for $\Lambda$ of the form $A = (\mathbf{a}_1 \ldots \mathbf{a}_r)$, which has rank $r$. In other words, a lattice $\Lambda$ of rank $r$ in $\mathbb{R}^n$ can always be described as $\Lambda = \mathbb{A} \mathbb{Z}^r$, where $A$ is its $m \times r$ basis matrix with real entries of rank $r$. As we remarked above, bases are not unique; as we will see later, each lattice has bases with particularly nice properties.

An important property of lattices is discreteness. To explain what we mean more notation is needed. First notice that Euclidean space $\mathbb{R}^n$ is clearly not compact, since it is not bounded. It is however locally compact: this means that for every point $\mathbf{x} \in \mathbb{R}^n$ there exists an open set containing $\mathbf{x}$ whose closure is compact, for instance take an open unit ball centered at $\mathbf{x}$. More generally, every subspace $V$ of $\mathbb{R}^n$ is also locally compact. A subset $\Gamma$ of $V$ is called discrete if for each $\mathbf{x} \in \Gamma$ there exists an open set $S \subseteq V$ such that $S \cap \Gamma = \{ \mathbf{x} \}$. For instance $\mathbb{Z}^n$ is a discrete subset of $\mathbb{R}^n$: for each point $\mathbf{x} \in \mathbb{Z}^n$ the open ball of radius $1/2$ centered at $\mathbf{x}$ contains no other points of $\mathbb{Z}^n$. We say that a discrete subset $\Gamma$ is co-compact
in $V$ if there exists a compact $0$-symmetric subset $U$ of $V$ such that the union of translations of $U$ by the points of $\Gamma$ covers the entire space $V$, i.e. if

$$V = \bigcup \{ U + x : x \in \Gamma \}.$$ 

Here $U + x = \{ u + x : u \in U \}$.

Recall that a subset $G$ is a subgroup of the additive abelian group $\mathbb{R}^n$ if it satisfies the following conditions:

1. **Identity:** $0 \in G$,
2. **Closure:** For every $x, y \in G$, $x + y \in G$,
3. **Inverses:** For every $x \in G$, $-x \in G$.

By Problems 1.3 and 1.4 a lattice $\Lambda$ of rank $r$ in $\mathbb{R}^n$ is a discrete co-compact subgroup of $V = \text{span}_\mathbb{R} \Lambda$. In fact, the converse is also true.

**Theorem 1.2.1.** Let $V$ be an $r$-dimensional subspace of $\mathbb{R}^n$, and let $\Gamma$ be a discrete co-compact subgroup of $V$. Then $\Gamma$ is a lattice of rank $r$ in $\mathbb{R}^n$.

**Proof.** In other words, we want to prove that $\Gamma$ has a basis, i.e. that there exists a collection of linearly independent vectors $a_1, \ldots, a_r$ in $\Gamma$ such that $\Gamma = \text{span}_\mathbb{Z} \{ a_1, \ldots, a_r \}$. We start by inductively constructing a collection of vectors $a_1, \ldots, a_r$, and then show that it has the required properties.

Let $a_1 \neq 0$ be a point in $\Gamma$ such that the line segment connecting $0$ and $a_1$ contains no other points of $\Gamma$. Now assume $a_1, \ldots, a_{i-1}, 2 \leq i \leq r$, have been selected; we want to select $a_i$. Let

$$H_{i-1} = \text{span}_\mathbb{R} \{ a_1, \ldots, a_{i-1} \},$$

and pick any $c \in \Gamma \setminus H_{i-1}$: such $c$ exists, since $\Gamma \not\subseteq H_{i-1}$ (otherwise $\Gamma$ would not be co-compact in $V$). Let $P_i$ be the closed parallelootope spanned by the vectors $a_1, \ldots, a_{i-1}, c$. Notice that since $\Gamma$ is discrete in $V$, $\Gamma \cap P_i$ is a finite set. Moreover, since $c \in P_i$, $\Gamma \cap P_i \not\subseteq H_{i-1}$. Then select $a_i$ such that

$$d(a_i, H_{i-1}) = \min_{y \in (P_i \cap \Gamma) \setminus H_{i-1}} \{ d(y, H_{i-1}) \},$$

where for any point $y \in \mathbb{R}^n$,

$$d(y, H_{i-1}) = \inf_{x \in H_{i-1}} \{ d(y, x) \}.$$

Let $a_1, \ldots, a_r$ be the collection of points chosen in this manner. Then we have

$$a_1 \neq 0, a_i \not\in \text{span}_\mathbb{Z} \{ a_1, \ldots, a_{i-1} \} \forall 2 \leq i \leq r,$$

which means that $a_1, \ldots, a_r$ are linearly independent. Clearly,

$$\text{span}_\mathbb{Z} \{ a_1, \ldots, a_r \} \subseteq \Gamma.$$

We will now show that

$$\Gamma \subseteq \text{span}_\mathbb{Z} \{ a_1, \ldots, a_r \}.$$

First of all notice that $a_1, \ldots, a_r$ is certainly a basis for $V$, and so if $x \in \Gamma \subseteq V$, then there exist $c_1, \ldots, c_r \in \mathbb{R}$ such that

$$x = \sum_{i=1}^r c_i a_i.$$
1. GEOMETRY OF NUMBERS

Notice that
\[ x' = \sum_{i=1}^{r} [c_i] a_i \in \text{span}_\mathbb{Z} \{a_1, \ldots, a_r\} \subseteq \Gamma, \]
where \([\ ]\) stands for the integer part function (i.e. \([c_i]\) is the largest integer which is no larger than \(c_i\)). Since \(\Gamma\) is a group, we must have
\[ z = x - x' = \sum_{i=1}^{r} (c_i - [c_i]) a_i \in \Gamma. \]
Then notice that
\[ d(z, H_{r-1}) = (c_r - [c_r]) d(a_r, H_{r-1}) < d(a_r, H_{r-1}), \]
but by construction we must have either \(z \in H_{r-1}\), or
\[ d(a_r, H_{r-1}) \leq d(z, H_{r-1}), \]
since \(z\) lies in the parallelotope spanned by \(a_1, \ldots, a_r\), and hence in \(P_r\) as in our construction above. Therefore \(c_r = [c_r]\). We proceed in the same manner to conclude that \(c_i = [c_i]\) for each \(1 \leq i \leq r\), and hence \(x \in \text{span}_\mathbb{Z} \{a_1, \ldots, a_r\}\). Since this is true for every \(x \in \Gamma\), we are done. \(\square\)

From now on, until further notice, our lattices will be of full rank in \(\mathbb{R}^n\), that is of rank \(n\). In other words, a lattice \(\Lambda \subseteq \mathbb{R}^n\) will be of the form \(\Lambda = AZ^n\), where \(A\) is a non-singular \(n \times n\) basis matrix for \(\Lambda\).

**Theorem 1.2.2.** Let \(\Lambda\) be a lattice of rank \(n\) in \(\mathbb{R}^n\), and let \(A\) be a basis matrix for \(\Lambda\). Then \(B\) is another basis matrix for \(\Lambda\) if and only if there exists an \(n \times n\) integral matrix \(U\) with determinant \(\pm 1\) such that
\[ B = AU. \]

**Proof.** First suppose that \(B\) is a basis matrix. Notice that, since \(A\) is a basis matrix, for every \(1 \leq i \leq n\) the \(i\)-th column vector \(b_i\) of \(B\) can be expressed as
\[ b_i = \sum_{j=1}^{n} u_{ij} a_j, \]
where \(a_1, \ldots, a_n\) are column vectors of \(A\), and \(u_{ij}\)’s are integers for all \(1 \leq j \leq n\). This means that \(B = AU\), where \(U = (u_{ij})_{1 \leq i, j \leq n}\) is an \(n \times n\) matrix with integer entries. On the other hand, since \(B\) is also a basis matrix, we also have for every \(1 \leq i \leq n\)
\[ a_i = \sum_{j=1}^{n} w_{ij} b_j, \]
where \(w_{ij}\)’s are also integers for all \(1 \leq j \leq N\). Hence \(A = BW\), where \(W = (w_{ij})_{1 \leq i, j \leq n}\) is also an \(n \times n\) matrix with integer entries. Then
\[ B = AU = BWU, \]
which means that \(WU = I_n\), the \(n \times n\) identity matrix. Therefore
\[ \det(WU) = \det(W) \det(U) = \det(I_n) = 1, \]
but \(\det(U), \det(W) \in \mathbb{Z}\) since \(U\) and \(W\) are integral matrices. This means that
\[ \det(U) = \det(W) = \pm 1. \]
Next assume that $B = UA$ for some integral $n \times n$ matrix $U$ with $\det(U) = \pm 1$. This means that $\det(B) = \pm \det(A) \neq 0$, hence column vectors of $B$ are linearly independent. Also, $U$ is invertible over $\mathbb{Z}$, meaning that $U^{-1} = (w_{ij})_{1 \leq i,j \leq n}$ is also an integral matrix, hence $A = U^{-1}B$. This means that column vectors of $A$ are in the span of the column vectors of $B$, and so

$$\Lambda \subseteq \text{span}_\mathbb{Z}\{b_1, \ldots, b_n\}.$$  

On the other hand, $b_i \in \Lambda$ for each $1 \leq i \leq n$. Thus $B$ is a basis matrix for $\Lambda$. □

**Corollary 1.2.3.** If $A$ and $B$ are two basis matrices for the same lattice $\Lambda$, then

$$|\det(A)| = |\det(B)|.$$  

**Definition 1.2.2.** The common determinant value of Corollary 1.2.3 is called the determinant of the lattice $\Lambda$, and is denoted by $\det(\Lambda)$.

We now talk about sublattices of a lattice. Let us start with a definition.

**Definition 1.2.3.** If $\Omega$ and $\Lambda$ are both lattices in $\mathbb{R}^n$, and $\Omega \subseteq \Lambda$, then we say that $\Omega$ is a sublattice of $\Lambda$.

There are a few basic properties of sublattices of a lattice which we outline here – their proofs are left to exercises.

1. A subset $\Omega$ of the lattice $\Lambda$ is a sublattice if and only if it is a subgroup of the abelian group $\Lambda$.
2. For a sublattice $\Omega$ of $\Lambda$ two cosets $x + \Omega$ and $y + \Omega$ are equal if and only if $x - y \in \Omega$. In particular, $x + \Omega = \Omega$ if and only if $x \in \Omega$.
3. If $\Lambda$ is a lattice and $\mu$ a real number, then the set

$$\mu\Lambda := \{\mu x : x \in \Lambda\}$$

is also a lattice. Further, if $\mu$ is an integer then $\mu\Lambda$ is a sublattice of $\Lambda$.

From here on, unless stated otherwise, when we say $\Omega \subseteq \Lambda$ is a sublattice, we always assume that it has the same full rank in $\mathbb{R}^n$ as $\Lambda$.

**Lemma 1.2.4.** Let $\Omega$ be a sublattice of $\Lambda$. There exists a positive integer $D$ such that $D\Lambda \subseteq \Omega$.

**Proof.** Recall that $\Lambda$ and $\Omega$ are both lattices of rank $n$ in $\mathbb{R}^n$. Let $a_1, \ldots, a_n$ be a basis for $\Omega$ and $b_1, \ldots, b_n$ be a basis for $\Lambda$. Then

$$\text{span}_\mathbb{R}\{a_1, \ldots, a_n\} = \text{span}_\mathbb{R}\{b_1, \ldots, b_n\} = \mathbb{R}^n.$$  

Since $\Omega \subseteq \Lambda$, there exist integers $u_{i1}, \ldots, u_{in}$ such that

$$\begin{cases} a_1 = u_{11} b_1 + \cdots + u_{1n} b_n \\ \vdots \\ a_n = u_{n1} b_1 + \cdots + u_{nn} b_n. \end{cases}$$

Solving this linear system for $b_1, \ldots, b_n$ in terms of $a_1, \ldots, a_n$, we easily see that there must exist rational numbers $\frac{p_{11}}{q_{11}}, \ldots, \frac{p_{nn}}{q_{nn}}$ such that

$$\begin{cases} b_1 = \frac{p_{11}}{q_{11}} a_1 + \cdots + \frac{p_{1n}}{q_{1n}} a_n \\ \vdots \\ b_n = \frac{p_{n1}}{q_{n1}} a_1 + \cdots + \frac{p_{nn}}{q_{nn}} a_n. \end{cases}$$
Let $D = q_1 \times \cdots \times q_n$, then $D/q_{ij} \in \mathbb{Z}$ for each $1 \leq i, j \leq n$, and so all the vectors
\[
\begin{cases}
Db_1 = \frac{Dp_{11}}{q_{11}} a_1 + \cdots + \frac{Dp_{1n}}{q_{nn}} a_n \\
\vdots \\
Db_n = \frac{Dp_{n1}}{q_{11}} a_1 + \cdots + \frac{Dp_{nn}}{q_{nn}} a_n
\end{cases}
\]
are in $\Omega$. Therefore $\text{span}_\mathbb{Z}\{Db_1, \ldots, Db_n\} \subseteq \Omega$. On the other hand,
\[
\text{span}_\mathbb{Z}\{Db_1, \ldots, Db_n\} = D\text{span}_\mathbb{Z}\{b_1, \ldots, b_n\} = D\Lambda,
\]
which completes the proof. 

We can now prove that a lattice always has a basis with “nice” properties with respect to any given basis of a given sublattice, and vice versa.

**Theorem 1.2.5.** Let $\Lambda$ be a lattice, and $\Omega$ a sublattice of $\Lambda$. For each basis $b_1, \ldots, b_n$ of $\Lambda$, there exists a basis $a_1, \ldots, a_n$ of $\Omega$ of the form
\[
\begin{cases}
a_1 = v_{11} b_1 \\
a_2 = v_{21} b_1 + v_{22} b_2 \\
\vdots \\
a_n = v_{n1} b_1 + \cdots + v_{nn} b_n,
\end{cases}
\]
where all $v_{ij} \in \mathbb{Z}$ and $v_{ii} \neq 0$ for all $1 \leq i \leq n$. Conversely, for every basis $a_1, \ldots, a_n$ of $\Omega$ there exists a basis $b_1, \ldots, b_n$ of $\Lambda$ such that the relations as above hold.

**Proof.** Let $b_1, \ldots, b_n$ be a basis for $\Lambda$. We will first prove the existence of a basis $a_1, \ldots, a_n$ for $\Omega$ as claimed by the theorem. By Lemma 1.2.4, there exist integer multiples of $b_1, \ldots, b_n$ in $\Omega$, hence it is possible to choose a collection of vectors $a_1, \ldots, a_n \in \Omega$ of the form
\[
a_i = \sum_{j=1}^{i} v_{ij} b_j,
\]
for each $1 \leq i \leq n$ with $v_{ii} \neq 0$. Clearly, by construction, such a collection of vectors will be linearly independent. In fact, let us pick each $a_i$ so that $|v_{ii}|$ is as small as possible, but not 0. We will now show that $a_1, \ldots, a_n$ is a basis for $\Omega$. Clearly, 
\[
\text{span}_\mathbb{Z}\{a_1, \ldots, a_n\} \subseteq \Omega.
\]

We want to prove the inclusion in the other direction, i.e. that
\[
\Omega \subseteq \text{span}_\mathbb{Z}\{a_1, \ldots, a_n\}.
\]
Suppose (1.1) is not true, then there exists $c \in \Omega$ which is not in $\text{span}_\mathbb{Z}\{a_1, \ldots, a_n\}$. Since $c \in \Lambda$, we can write
\[
c = \sum_{j=1}^{k} t_j b_j,
\]
for some integers $1 \leq k \leq n$ and $t_1, \ldots, t_k$. In fact, let us select $c$ like this with minimal possible $k$. Since $v_{kk} \neq 0$, we can choose an integer $s$ such that
\[
|t_k - sv_{kk}| < |v_{kk}|.
\]
Then we clearly have 
\[
c - sa_k \in \Omega \setminus \text{span}_\mathbb{Z}\{a_1, \ldots, a_n\}.
\]
Therefore we must have \( t_k - sv_{kk} \neq 0 \) by minimality of \( k \). But then (1.2) contradicts the minimality of \( |v_{kk}| \): we could take \( c - sa_k \) instead of \( a_k \), since it satisfies all the conditions that \( a_k \) was chosen to satisfy, and then \( |v_{kk}| \) is replaced by the smaller nonzero number \( |t_k - sv_{kk}| \). This proves that \( c \) like this cannot exist, and so (1.1) is true, hence finishing one direction of the theorem.

Now suppose that we are given a basis \( a_1, \ldots, a_n \) for \( \Omega \). We want to prove that there exists a basis \( b_1, \ldots, b_n \) for \( \Lambda \) such that relations in the statement of the theorem hold. This is a direct consequence of the argument in the proof of Theorem 1.2.1. Indeed, at \( i \)-th step of the basis construction in the proof of Theorem 1.2.1, we can choose \( i \)-th vector, call it \( b_i \), so that it lies in the span of the previous \( i - 1 \) vectors and the vector \( a_i \). Since \( b_1, \ldots, b_n \) constructed this way are linearly independent (in fact, they form a basis for \( \Lambda \) by the construction), we obtain that

\[
a_i \in \text{span}\{b_1, \ldots, b_i\} \setminus \text{span}\{b_1, \ldots, b_{i-1}\},
\]

for each \( 1 \leq i \leq n \). This proves the second half of our theorem.

In fact, it is possible to select the coefficients \( v_{ij} \) in Theorem 1.2.5 so that the matrix \( (v_{ij})_{1 \leq i,j \leq n} \) is upper (or lower) triangular with non-negative entries, and the largest entry of each row (or column) is on the diagonal: we leave the proof of this to Problem 1.9.

Remark 1.2.1. Let the notation be as in Theorem 1.2.5. Notice that if \( A \) is any basis matrix for \( \Omega \) and \( B \) is any basis for \( \Lambda \), then there exists an integral matrix \( V \) such that \( A = BV \). Then Theorem 1.2.5 implies that for a given \( B \) there exists an \( A \) such that \( V \) is lower triangular, and for some \( A \) there exists a \( B \) such that \( V \) is lower triangular. Since two different basis matrices of the same lattice are always related by multiplication by an integral matrix with determinant equal to \( \pm 1 \), Theorem 1.2.5 can be thought of as the construction of Hermite normal form for an integral matrix. Problem 1.9 places additional restrictions that make Hermite normal form unique.

Here is an important implication of Theorem 1.2.5.

Theorem 1.2.6. Let \( \Omega \subseteq \Lambda \) be a sublattice. Then \( \frac{\det(\Omega)}{\det(\Lambda)} \) is an integer; moreover, the number of cosets of \( \Omega \) in \( \Lambda \), i.e. the index of \( \Omega \) as a subgroup of \( \Lambda \) is

\[
[\Lambda : \Omega] = \frac{\det(\Omega)}{\det(\Lambda)}.
\]

Proof. Let \( b_1, \ldots, b_n \) be a basis for \( \Lambda \), and \( a_1, \ldots, a_n \) be a basis for \( \Omega \), so that these two bases satisfy the conditions of Theorem 1.2.5, and write \( A \) and \( B \) for the corresponding basis matrices. Then notice that

\[
B = AV,
\]

where \( V = (v_{ij})_{1 \leq i,j \leq n} \) is an \( n \times n \) triangular matrix with entries as described in Theorem 1.2.5; in particular \( \det(V) = \prod_{i=1}^{n} |v_{ii}|. \) Hence

\[
\det(\Omega) = |\det(A)| = |\det(B)||\det(V)| = \det(\Lambda) \prod_{i=1}^{n} |v_{ii}|.
\]

which proves the first part of the theorem.
Moreover, notice that each vector \( c \in \Lambda \) is contained in the same coset of \( \Omega \) in \( \Lambda \) as precisely one of the vectors \( q_1 b_1 + \cdots + q_n b_n \), \( 0 \leq q_i < v_{ii} \forall 1 \leq i \leq n \), in other words there are precisely \( \prod_{i=1}^{n} |v_{ii}| \) cosets of \( \Omega \) in \( \Lambda \). This completes the proof. \( \square \)

There is yet another, more analytic interpretation of the determinant of a lattice.

**Definition 1.2.4.** A fundamental domain of a lattice \( \Lambda \) of full rank in \( \mathbb{R}^n \) is a convex set \( F \subseteq \mathbb{R}^n \) containing \( 0 \), so that \( \mathbb{R}^n = \bigcup_{x \in \Lambda} (F + x) \), and for every \( x \neq y \in \Lambda \), \( (F + x) \cap (F + y) = \emptyset \).

In other words, a fundamental domain of a lattice \( \Lambda \subseteq \mathbb{R}^n \) is a full set of coset representatives of \( \Lambda \) in \( \mathbb{R}^n \) (see Problem 1.10). Although each lattice has infinitely many different fundamental domains, they all have the same volume, which is equal to the determinant of the lattice. This fact can be easily proved for a special class of fundamental domains (see Problem 1.11).

**Definition 1.2.5.** Let \( \Lambda \) be a lattice, and \( a_1, \ldots, a_n \) be a basis for \( \Lambda \). Then the set

\[
F = \left\{ \sum_{i=1}^{n} t_i a_i : 0 \leq t_i < 1, \quad \forall 1 \leq i \leq n \right\},
\]

is called a fundamental parallelotope of \( \Lambda \) with respect to the basis \( a_1, \ldots, a_n \). It is easy to see that this is an example of a fundamental domain for a lattice.

Fundamental parallelotopes form the most important class of fundamental domains, which we will work with most often. Notice that they are not closed sets; we will often write \( \overline{F} \) for the closure of a fundamental parallelotope, and call them closed fundamental domains. Another important convex set associated to a lattice is its Voronoi cell, which is the closure of a fundamental domain; by a certain abuse of notation we will often refer to it also as a fundamental domain.

**Definition 1.2.6.** The Voronoi cell of a lattice \( \Lambda \) is the set

\[
V(\Lambda) = \{ x \in \mathbb{R}^n : \|x\| \leq \|x - y\| \quad \forall \ y \in \Lambda \}.
\]

It is easy to see that \( V(\Lambda) \) is (the closure of) a fundamental domain for \( \Lambda \): two translates of a Voronoi cell by points of the lattice intersect only in the boundary. The advantage of the Voronoi cell is that it is the most "round" fundamental domain for a lattice; we will see that it comes up very naturally in the context of sphere packing and covering problems.

Notice that everything we discussed so far also has analogues for lattices of not necessarily full rank. We mention this here briefly without proofs. Let \( \Lambda \) be a lattice in \( \mathbb{R}^n \) of rank \( 1 \leq r \leq n \), and let \( a_1, \ldots, a_r \) be a basis for it. Write \( A = (a_1 \ldots a_r) \) for the corresponding \( n \times r \) basis matrix of \( \Lambda \), then \( A \) has rank \( r \) since its column vectors are linearly independent. For any \( r \times r \) integral matrix \( U \) with determinant \( \pm 1 \), \( AU \) is another basis matrix for \( \Lambda \); moreover, if \( B \) is any other basis matrix for
Let $\Lambda$, there exists such a $U$ so that $B = AU$. For each basis matrix $A$ of $\Lambda$, we define the corresponding Gram matrix to be $M = A^T A$, so it is a square $r \times r$ nonsingular matrix. Notice that if $A$ and $B$ are two basis matrices so that $B = UA$ for some $U$ as above, then

$$
\det(B^T B) = \det((AU)^T (AU)) = \det(U^T (A^T A) U) = \det(U)^2 \det(A^T A) = \det(A^T A).
$$

This observation calls for the following general definition of the determinant of a lattice. Notice that this definition coincides with the previously given one in case $r = n$.

**Definition 1.2.7.** Let $\Lambda$ be a lattice of rank $1 \leq r \leq n$ in $\mathbb{R}^n$, and let $A$ be an $n \times r$ basis matrix for $\Lambda$. The determinant of $\Lambda$ is defined to be

$$
\det(\Lambda) = \sqrt{\det(A^T A)},
$$

that is the determinant of the corresponding Gram matrix. By the discussion above, this is well defined, i.e. does not depend on the choice of the basis.

With this notation, all results and definitions of this section can be restated for a lattice $\Lambda$ of not necessarily full rank. For instance, in order to define fundamental domains we can view $\Lambda$ as a lattice inside of the vector space $\text{span}_\mathbb{R}(\Lambda)$. The rest works essentially verbatim, keeping in mind that if $\Omega \subseteq \Lambda$ is a sublattice, then index $[\Lambda : \Omega]$ is only defined if $\text{rk}(\Omega) = \text{rk}(\Lambda)$. 

1. Theorems of Blichfeldt and Minkowski

In this section we will discuss some of the famous theorems related to the following very classical problem in the geometry of numbers: given a set $M$ and a lattice $\Lambda$ in $\mathbb{R}^n$, how can we tell if $M$ contains any points of $\Lambda$?

**Theorem 1.3.1 (Blichfeldt, 1914).** Let $M$ be a compact convex set in $\mathbb{R}^n$. Suppose that $\text{Vol}(M) \geq 1$. Then there exist $x, y \in M$ such that $0 \neq x - y \in \mathbb{Z}^n$.

**Proof.** First suppose that $\text{Vol}(M) > 1$. Let $P = \{ x \in \mathbb{R}^n : 0 \leq x_i < 1 \ \forall \ 1 \leq i \leq n \}$, and let $S = \{ u \in \mathbb{Z}^n : M \cap (P + u) \neq \emptyset \}$. Since $M$ is bounded, $S$ is a finite set, say $S = \{ u_1, \ldots, u_{r_0} \}$. Write $M_r = M \cap (P + u_r)$ for each $1 \leq r \leq r_0$. Also, for each $1 \leq r \leq r_0$, define $M'_r = M_r - u_r$, so that $M'_1, \ldots, M'_{r_0}, M,r_0 \subseteq P$. On the other hand, $\bigcup_{r=1}^{r_0} M_r = M$, and $M_r \cap M_s = \emptyset$ for all $1 \leq r \neq s \leq r_0$, since $M_r \subseteq P + u_r, M_s \subseteq P + u_s$, and $(P + u_r) \cap (P + u_s) = \emptyset$. This means that

$$1 < \text{Vol}(M) = \sum_{r=1}^{r_0} \text{Vol}(M_r).$$

However, $\text{Vol}(M'_r) = \text{Vol}(M_r)$ for each $1 \leq r \leq r_0$,

$$\sum_{r=1}^{r_0} \text{Vol}(M'_r) > 1,$$

but $\bigcup_{r=1}^{r_0} M'_r \subseteq P$, and so

$$\text{Vol} \left( \bigcup_{r=1}^{r_0} M'_r \right) \leq \text{Vol}(P) = 1.$$

Hence the sets $M'_1, \ldots, M'_{r_0}$ are not mutually disjoined, meaning that there exist indices $1 \leq r \neq s \leq r_0$ such that there exists $x \in M'_r \cap M'_s$. Then we have $x + u_r, x + u_s \in M$, and

$$(x + u_r) - (x + u_s) = u_r - u_s \in \mathbb{Z}^n.$$ 

Now suppose $M$ is closed, bounded, and $\text{Vol}(M) = 1$. Let $\{ s_r \}_{r=1}^{\infty}$ be a sequence of numbers all greater than 1, such that

$$\lim_{r \to \infty} s_r = 1.$$

By the argument above we know that for each $r$ there exist $x_r \neq y_r \in s_r M$ such that $x_r - y_r \in \mathbb{Z}^n$. Then there are subsequences $\{ x_{r_k} \}$ and $\{ y_{r_k} \}$ converging to points $x, y \in M$, respectively. Since for each $r_k$, $x_{r_k} - y_{r_k}$ is a nonzero lattice point, it must be true that $x \neq y$, and $x - y \in \mathbb{Z}^n$. This completes the proof. □

As a corollary of Theorem 1.3.1 we can prove the following version of *Minkowski Convex Body Theorem.*
Theorem 1.3.2 (Minkowski). Let $M \subset \mathbb{R}^n$ be a compact convex $0$-symmetric set with $\text{Vol}(M) \geq 2^n$. Then there exists $0 \neq x \in M \cap \mathbb{Z}^n$.

**Proof.** Notice that the set 
\[ \frac{1}{2} M = \left\{ \frac{1}{2} x : x \in M \right\} = \left( \begin{array}{ccc} 1/2 & 0 & \ldots & 0 \\ 0 & 1/2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/2 \end{array} \right) M \]

is also convex, $0$-symmetric, and by Problem 1.12 its volume is 
\[ \text{det} \left( \begin{array}{ccc} 1/2 & 0 & \ldots & 0 \\ 0 & 1/2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/2 \end{array} \right) \text{Vol}(M) = 2^{-n} \text{Vol}(M) \geq 1. \]

Therefore, by Theorem 1.3.1, there exist $\frac{1}{2} x \neq \frac{1}{2} y \in \frac{1}{2} M$ such that 
\[ \frac{1}{2} x - \frac{1}{2} y \in \mathbb{Z}^n. \]

But, by symmetry, since $y \in M$, $-y \in M$, and by convexity, since $x, -y \in M$, 
\[ \frac{1}{2} x - \frac{1}{2} y = \frac{1}{2} x + \frac{1}{2} (-y) \in M. \]

This completes the proof. \qed

Remark 1.3.1. This result is sharp: for any $\varepsilon > 0$, the cube 
\[ C = \left\{ x \in \mathbb{R}^n : \max_{1 \leq i \leq n} |x_i| \leq 1 - \varepsilon \right\} \]

is a convex $0$-symmetric set of volume $(2-\varepsilon)^n$, which contains no nonzero integer lattice points.

Problem 1.13 extends Blichfeldt and Minkowski theorems to arbitrary lattices as follows:

- If $\Lambda \subset \mathbb{R}^n$ is a lattice of full rank and $M \subset \mathbb{R}^n$ is a compact convex set with $\text{Vol}(M) \geq \text{det} \Lambda$, then there exist $x, y \in M$ such that $0 \neq x - y \in \Lambda$.
- If $\Lambda \subset \mathbb{R}^n$ is a lattice of full rank and $M \subset \mathbb{R}^n$ is a compact convex $0$-symmetric set with $\text{Vol}(M) \geq 2^n \text{det} \Lambda$, then there exists $0 \neq x \in M \cap \Lambda$.

As a first application of these results, we now prove Minkowski’s Linear Forms Theorem.

**Theorem 1.3.3.** Let $B = (b_{ij})_{1 \leq i, j \leq n} \in \text{GL}_n(\mathbb{R})$, and for each $1 \leq i \leq n$ define a linear form with coefficients $b_{i1}, \ldots, b_{in}$ by 
\[ L_i(X) = \sum_{j=1}^{n} b_{ij} X_j. \]

Let $c_1, \ldots, c_n \in \mathbb{R}_{>0}$ be such that 
\[ c_1 \cdots c_n = |\text{det}(B)|. \]

Then there exists $0 \neq x \in \mathbb{Z}^n$ such that 
\[ |L_i(x)| \leq c_i, \]
for each $1 \leq i \leq n$.

**Proof.** Let us write $b_1, \ldots, b_n$ for the row vectors of $B$, then

$$L_i(x) = b_i x,$$

for each $x \in \mathbb{R}^n$. Consider parallelepiped

$$P = \{ x \in \mathbb{R}^n : |L_i(x)| \leq c_i \ \forall \ 1 \leq i \leq n \} = B^{-1} R,$$

where $R = \{ x \in \mathbb{R}^n : |x_i| \leq c_i \ \forall \ 1 \leq i \leq n \}$ is the rectangular box with sides of length $2c_1, \ldots, 2c_n$ centered at the origin in $\mathbb{R}^n$. Then by Problem 1.12,

$$\text{Vol}(P) = |\det(B)|^{-1} \text{Vol}(R) = |\det(B)|^{-1} 2^n c_1 \cdots c_n = 2^n,$$

and so by Theorem 1.3.2 there exists $0 \neq x \in P \cap \mathbb{Z}^n$. \qed
1.4. Successive minima

Let us start with a certain restatement of Minkowski’s Convex Body theorem.

Corollary 1.4.1. Let \( M \subseteq \mathbb{R}^n \) be a compact convex 0-symmetric and \( \Lambda \subseteq \mathbb{R}^n \) a lattice of full rank. Define the first successive minimum of \( M \) with respect to \( \Lambda \) to be

\[
\lambda_1 = \inf \{ \lambda \in \mathbb{R}_{>0} : \lambda M \cap \Lambda \text{ contains a nonzero point} \}.
\]

Then

\[
0 < \lambda_1 \leq 2^{n/n} \left( \frac{\det \Lambda}{\text{Vol}(M)} \right)^{1/n}.
\]

Proof. The fact that \( \lambda_1 \) has to be positive readily follows from \( \Lambda \) being a discrete set. Hence we only have to prove the upper bound. By Theorem 1.3.2 for a general lattice \( \Lambda \) (Problem 1.13), if

\[
\text{Vol}(\lambda M) \geq 2^n \det(\Lambda),
\]

then \( \lambda M \) contains a nonzero point of \( \Lambda \). On the other hand, by Problem 1.12,

\[
\text{Vol}(\lambda M) = \lambda^n \text{Vol}(M).
\]

Hence as long as

\[
\lambda^n \text{Vol}(M) \geq 2^n \det(\Lambda),
\]

the expanded set \( \lambda M \) is guaranteed to contain a nonzero point of \( \Lambda \). The conclusion of the corollary follows. \( \square \)

The above corollary thus provides an estimate as to how much should the set \( M \) be expanded to contain a nonzero point of the lattice \( \Lambda \): this is the meaning of \( \lambda_1 \), it is precisely this expansion factor. A natural next question to ask is how much should we expand \( M \) to contain 2 linearly independent points of \( \Lambda \), 3 linearly independent points of \( \Lambda \), etc. To answer this question is the main objective of this section. We start with a definition.

Definition 1.4.1. Let \( M \) be a convex, 0-symmetric set \( M \subseteq \mathbb{R}^n \) of non-zero volume and \( \Lambda \subseteq \mathbb{R}^n \) a lattice of full rank. For each \( 1 \leq i \leq n \) define the \( i \)-th successive minimum of \( M \) with respect to \( \Lambda \), \( \lambda_i \), to be the infimum of all positive real numbers \( \lambda \) such that the set \( \lambda M \) contains at least \( i \) linearly independent points of \( \Lambda \). In other words,

\[
\lambda_i = \inf \{ \lambda \in \mathbb{R}_{>0} : \dim (\text{span}_\mathbb{R}\{\lambda M \cap \Lambda\}) \geq i \}.
\]

Since \( \Lambda \) is discrete in \( \mathbb{R}^n \), the infimum in this definition is always achieved, i.e. it is actually a minimum.

Remark 1.4.1. Notice that the \( n \) linearly independent vectors \( \mathbf{u}_1, \ldots, \mathbf{u}_n \) corresponding to successive minima \( \lambda_1, \ldots, \lambda_n \), respectively, do not necessarily form a basis. It was already known to Minkowski that they do in dimensions \( n = 1, \ldots, 4 \), but when \( n = 5 \) there is a well known counterexample. Let

\[
\Lambda = \begin{pmatrix}
1 & 0 & 0 & 0 & \frac{1}{2} \\
0 & 1 & 0 & 0 & \frac{1}{2} \\
0 & 0 & 1 & 0 & \frac{1}{2} \\
0 & 0 & 0 & 1 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 1
\end{pmatrix} \subseteq \mathbb{Z}^5,
\]
and let $M = B_5$, the closed unit ball centered at $0$ in $\mathbb{R}^n$. Then the successive minima of $B_5$ with respect to $\Lambda$ is

$$\lambda_1 = \cdots = \lambda_5 = 1,$$

since $e_1, \ldots, e_5 \in B_5 \cap \Lambda$, and

$$x = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix}^\top \notin B_5.$$

On the other hand, $x$ cannot be expressed as a linear combination of $e_1, \ldots, e_5$ with integer coefficients, hence

$$\text{span}_\mathbb{Z}\{e_1, \ldots, e_5\} \subset \Lambda.$$

An immediate observation is that

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

and Corollary 1.4.1 gives an upper bound on $\lambda_1$. Can we produce bounds on all the successive minima in terms of $\text{Vol}(M)$ and $\det(\Lambda)$? This question is answered by Minkowski’s Successive Minima Theorem.

**Theorem 1.4.2.** With notation as above,

$$\frac{2^n \det(\Lambda)}{n! \text{Vol}(M)} \leq \lambda_1 \cdots \lambda_n \leq \frac{2^n \det(\Lambda)}{\text{Vol}(M)}.$$

**Proof.** We present the proof in case $\Lambda = \mathbb{Z}^n$, leaving generalization of the given argument to arbitrary lattices as an exercise. We start with a proof of the lower bound following [GL87], which is considerably easier than the upper bound. Let $u_1, \ldots, u_n$ be the $n$ linearly independent vectors corresponding to the respective successive minima $\lambda_1, \ldots, \lambda_n$, and let

$$U = (u_1 \ldots u_n) = \begin{pmatrix} u_{11} & \cdots & u_{n1} \\ \vdots & \ddots & \vdots \\ u_{1n} & \cdots & u_{nn} \end{pmatrix}.$$

Then $U = U \mathbb{Z}^n$ is a full rank sublattice of $\mathbb{Z}^n$ with index $|\det(U)|$. Notice that the $2n$ points

$$\pm \frac{u_1}{\lambda_1}, \ldots, \pm \frac{u_n}{\lambda_n}$$

lie in $M$, hence $M$ contains the convex hull $P$ of these points, which is a generalized octahedron. Any polyhedron in $\mathbb{R}^n$ can be decomposed as a union of simplices that pairwise intersect only in the boundary. A standard simplex in $\mathbb{R}^n$ is the convex hull of $n$ points, so that no 3 of them are co-linear, no 4 of them are co-planar, etc., no $k$ of them lie in a $(k-1)$-dimensional subspace of $\mathbb{R}^n$, and so that their convex hull does not contain any integer lattice points in its interior. The volume of a standard simplex in $\mathbb{R}^n$ is $1/n!$ (Problem 1.14).

Our generalized octahedron $P$ can be decomposed into $2^n$ simplices, which are obtained from the standard simplex by multiplication by the matrix

$$\begin{pmatrix} \frac{u_{11}}{\lambda_1} & \cdots & \frac{u_{n1}}{\lambda_n} \\ \vdots & \ddots & \vdots \\ \frac{u_{1n}}{\lambda_1} & \cdots & \frac{u_{nn}}{\lambda_n} \end{pmatrix}.$$
therefore its volume is

$$\text{(1.3)} \quad \text{Vol}(P) = \frac{2^n}{n!} \left| \det \begin{pmatrix} u_{11} & \cdots & u_{1N} \\ \lambda_1 & \cdots & \lambda_N \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nN} \end{pmatrix} \right| = \frac{2^n |\det(U)|}{n! \lambda_1 \cdots \lambda_N} \geq \frac{2^n}{n! \lambda_1 \cdots \lambda_N},$$

since $\det(U)$ is an integer. Since $P \subseteq M$, $\text{Vol}(M) \geq \text{Vol}(P)$. Combining this last observation with (1.3) yields the lower bound of the theorem.

Next we prove the upper bound. The argument we present is due to M. Henk [Hen02], and is at least partially based on Minkowski’s original geometric ideas. For each $1 \leq i \leq n$, let

$$E_i = \text{span}_\mathbb{R}\{e_1, \ldots, e_i\},$$

the $i$-th coordinate subspace of $\mathbb{R}^n$, and define

$$M_i = \frac{\lambda_i}{2} M.$$

As in the proof of the lower bound, we take $u_1, \ldots, u_n$ to be the $n$ linearly independent vectors corresponding to the respective successive minima $\lambda_1, \ldots, \lambda_n$. In fact, notice that there exists a matrix $A \in \text{GL}_n(\mathbb{Z})$ such that

$$A\text{span}_\mathbb{R}\{u_1, \ldots, u_i\} \subseteq E_i,$$

for each $1 \leq i \leq n$, i.e. we can rotate each $\text{span}_\mathbb{R}\{u_1, \ldots, u_i\}$ so that it is contained in $E_i$. Moreover, volume of $AM$ is the same as volume of $M$, since $\det(A) = 1$ (i.e. rotation does not change volumes), and

$$Au_i \in \lambda'_i AM \cap E_i, \ \forall \ 1 \leq i \leq n,$$

where $\lambda'_1, \ldots, \lambda'_n$ is the successive minima of $AM$ with respect to $\mathbb{Z}^n$. Hence we can assume without loss of generality that

$$\text{span}_\mathbb{R}\{u_1, \ldots, u_i\} \subseteq E_i,$$

for each $1 \leq i \leq n$.

For an integer $q \in \mathbb{Z}_{>0}$, define the integral cube of sidelength $2q$ centered at $0$ in $\mathbb{R}^n$

$$C^n_q = \{z \in \mathbb{Z}^n : |z| \leq q\},$$

and for each $1 \leq i \leq n$ define the section of $C^n_q$ by $E_i$

$$C^i_q = C^n_q \cap E_i.$$

Notice that $C^n_q$ is contained in real cube of volume $(2q)^n$, and so the volume of all translates of $M$ by the points of $C^n_q$ can be bounded

$$\text{(1.4)} \quad \text{Vol}(C^n_q + M_n) \leq (2q + \gamma)^n,$$

where $\gamma$ is a constant that depends on $M$ only. Also notice that if $x \neq y \in \mathbb{Z}^n$, then

$$\text{int}(x + M_1) \cap \text{int}(y + M_1) = \emptyset,$$

where int stands for interior of a set: suppose not, then there exists

$$z \in \text{int}(x + M_1) \cap \text{int}(y + M_1),$$
and so
\[(z - x) - (z - y) = y - x \in \text{int}(M_1) - \text{int}(M_1)\]
(1.5) \[\{z_1 - z_2 : z_1, z_2 \in M_1\} = \text{int}(\lambda_1 M),\]
which would contradict minimality of \(\lambda_1\). Therefore
\[(1.6) \quad \text{Vol}(C^n_q + M_1) = (2q + 1)^n \text{Vol}(M_1) = (2q + 1)^n \left(\frac{\lambda_1}{2}\right)^n \text{Vol}(M).\]

To finish the proof, we need the following lemma.

**Lemma 1.4.3.** For each \(1 \leq i \leq n - 1\),
\[(1.7) \quad \text{Vol}(C^n_q + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i} \text{Vol}(C^n_q + M_i).\]

**Proof.** If \(\lambda_{i+1} = \lambda_i\) the statement is obvious, so assume \(\lambda_{i+1} > \lambda_i\). Let \(x, y \in \mathbb{Z}^n\) be such that
\[(x_{i+1}, \ldots, x_n) \neq (y_{i+1}, \ldots, y_n).\]
Then
\[(1.8) \quad (x + \text{int}(M_{i+1})) \cap (y + \text{int}(M_{i+1})) = \emptyset.\]
Indeed, suppose (1.8) is not true, i.e. there exists \(z \in (x + \text{int}(M_{i+1})) \cap (y + \text{int}(M_{i+1}))\). Then, as in (1.5) above, \(x - y \in \text{int}(\lambda_{i+1} M)\). But we also have
\[u_1, \ldots, u_i \in \text{int}(\lambda_{i+1} M),\]
since \(\lambda_{i+1} > \lambda_i\), and so \(\lambda_i M \subseteq \text{int}(\lambda_{i+1} M)\). Moreover, \(u_1, \ldots, u_i \in E_i\), meaning that
\[u_{jk} = 0 \quad \forall \quad 1 \leq j \leq i, \quad i + 1 \leq k \leq n.\]
On the other hand, at least one of
\[x_k - y_k, \quad i + 1 \leq k \leq n,\]
is not equal to 0. Hence \(x - y, u_{1,1}, \ldots, u_i\) are linearly independent, but this means that \(\text{int}(\lambda_{i+1} M)\) contains \(i + 1\) linearly independent points, contradicting minimality of \(\lambda_{i+1}\). This proves (1.8). Notice that (1.8) implies
\[\text{Vol}(C^n_q + M_{i+1}) = (2q + 1)^n \text{Vol}(C^n_q + M_{i+1}),\]
and
\[\text{Vol}(C^n_q + M_i) = (2q + 1)^{n-i} \text{Vol}(C^n_q + M_i),\]
since \(M_i \subseteq M_{i+1}\). Hence, in order to prove the lemma it is sufficient to prove that
\[(1.9) \quad \text{Vol}(C^n_q + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i} \text{Vol}(C^n_q + M_i).\]

Define two linear maps \(f_1, f_2 : \mathbb{R}^n \to \mathbb{R}^n\), given by
\[
f_1(x) = \left(\frac{\lambda_{i+1}}{\lambda_i} x_1, \ldots, \frac{\lambda_{i+1}}{\lambda_i} x_i, x_{i+1}, \ldots, x_n\right),
\]
\[
f_2(x) = \left(x_1, \ldots, x_i, \frac{\lambda_{i+1}}{\lambda_i} x_{i+1}, \ldots, \frac{\lambda_{i+1}}{\lambda_i} x_n\right),
\]
and notice that \(f_2(f_1(M_i)) = M_{i+1}\), \(f_2(C^n_q) = C^n_q\). Therefore
\[f_2(C^n_q + f_1(M_i)) = C^n_q + f_1(M_{i+1}).\]
This implies that
\[ \text{Vol}(C_q^i + M_{i+1}) = \left( \frac{\lambda_{i+1}}{\lambda_i} \right)^{n-i} \text{Vol}(C_q^i + f_i(M_i)), \]
and so to establish (1.9) it is sufficient to show that
\[ (1.10) \quad \text{Vol}(C_q^i + f_1(M_i)) \geq \text{Vol}(C_q^i + M_i). \]
Let
\[ E_i^+ = \text{span}_\mathbb{R} \{ e_{i+1}, \ldots, e_n \}, \]
i.e. \( E_i^+ \) is the orthogonal complement of \( E_i \), and so has dimension \( n - i \). Notice that for every \( x \in E_i^+ \) there exists \( t(x) \in E_i \) such that \( M_i \cap (x + E_i) \subseteq (f_1(M_i) \cap (x + E_i)) + t(x) \), in other words, although it is not necessarily true that \( M_i \subseteq f_1(M_i) \), each section of \( M_i \) by a translate of \( E_i \) is contained in a translate of some such section of \( f_1(M_i) \). Therefore
\[ (C_q^i + M_i) \cap (x + E_i) \subseteq (C_q^i + f_1(M_i)) \cap (x + E_i) + t(x), \]
and hence
\[
\text{Vol}(C_q^i + M_i) = \int_{x \in E_i^+} \text{Vol}_i((C_q^i + M_i) \cap (x + E_i)) \, dx \\
\leq \int_{x \in E_i^+} \text{Vol}_i((C_q^i + f_1(M_i)) \cap (x + E_i)) \, dx \\
= \text{Vol}(C_q^i + f_1(M_i)),
\]
where \( \text{Vol}_i \) stands for the \( i \)-dimensional volume. This completes the proof of (1.10), and hence of the lemma. \( \square \)

Now, combining (1.4), (1.6), and (1.7), we obtain:
\[
(2q + \gamma)^n \geq \text{Vol}(C_q^n + M_n) \geq \left( \frac{\lambda_n}{\lambda_{n-1}} \right) \text{Vol}(C_q^n + M_{n-1}) \geq \cdots \\
\geq \left( \frac{\lambda_n}{\lambda_{n-1}} \right) \left( \frac{\lambda_{n-1}}{\lambda_{n-2}} \right)^2 \cdots \left( \frac{\lambda_2}{\lambda_1} \right)^{n-1} \text{Vol}(C_q^n + M_1) \\
= \lambda_n \cdots \lambda_1 \frac{\text{Vol}(M)}{2^n} (2q + 1)^n,
\]
hence
\[
\lambda_1 \cdots \lambda_n \leq \frac{2^n}{\text{Vol}(M)} \left( \frac{2q + \gamma}{2q + 1} \right)^n \xrightarrow{q \to \infty} \frac{2^n}{\text{Vol}(M)},
\]
as \( q \to \infty \), since \( q \in \mathbb{Z}_{>0} \) is arbitrary. This completes the proof. \( \square \)

We can talk about successive minima of any convex \( 0 \)-symmetric set in \( \mathbb{R}^n \) with respect to the lattice \( \Lambda \). Perhaps the most frequently encountered such set is the closed unit ball \( \mathbb{B}_n \) in \( \mathbb{R}^n \) centered at \( 0 \). We define the \textit{successive minima of} \( \Lambda \) to be the successive minima of \( \mathbb{B}_n \) with respect to \( \Lambda \). Notice that successive minima are invariants of the lattice.
1.5. Inhomogeneous minimum

Here we exhibit one important application of Minkowski’s successive minima theorem. As before, let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank, and let $M \subseteq \mathbb{R}^n$ be a convex 0-symmetric set of nonzero volume. Throughout this section, we let
\[
\lambda_1 \leq \cdots \leq \lambda_n
\]
to be the successive minima of $M$ with respect to $\Lambda$. We define the inhomogeneous minimum of $M$ with respect to $\Lambda$ to be
\[
\mu = \inf \{ \lambda \in \mathbb{R}_{>0} : \lambda M + \Lambda = \mathbb{R}^n \}.
\]

The main objective of this section is to obtain some basic bounds on $\mu$. We start with the following result of Jarnik [Jar41].

**Lemma 1.5.1.**
\[
\mu \leq \frac{1}{2} \sum_{i=1}^{n} \lambda_i.
\]

**Proof.** Let us define a function
\[
F(x) = \inf \{ a \in \mathbb{R}_{>0} : x \in aM \},
\]
for every $x \in \mathbb{R}^n$. This function is a norm (Problem 1.15). Then
\[
M = \{ x \in \mathbb{R}^n : F(x) \leq 1 \}
\]
can be thought of as the unit ball with respect to this norm. We will say that $F$ is the norm of $M$. Let $z \in \mathbb{R}^n$ be an arbitrary point. We want to prove that there exists a point $v \in \Lambda$ such that
\[
F(z - v) \leq \frac{1}{2} \sum_{i=1}^{n} \lambda_i.
\]
This would imply that $z \in \left( \frac{1}{2} \sum_{i=1}^{n} \lambda_i \right) M + v$, and hence settle the lemma, since $z$ is arbitrary. Let $u_1, \ldots, u_n$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \ldots, \lambda_n$, respectively. Then
\[
F(u_i) = \lambda_i, \quad \forall \ 1 \leq i \leq n.
\]
Since $u_1, \ldots, u_n$ form a basis for $\mathbb{R}^n$, there exist $a_1, \ldots, a_n \in \mathbb{R}$ such that
\[
z = \sum_{i=1}^{n} a_i u_i.
\]
We can also choose integer $v_1, \ldots, v_n$ such that
\[
|a_i - v_i| \leq \frac{1}{2} , \quad \forall \ 1 \leq i \leq n,
\]
and define $v = \sum_{i=1}^{n} v_i u_i$, hence $v \in \Lambda$. Now notice that
\[
F(z - v) = F\left( \sum_{i=1}^{n} (a_i - v_i) u_i \right)
\]
\[
\leq \sum_{i=1}^{n} |a_i - v_i| F(u_i) \leq \frac{1}{2} \sum_{i=1}^{n} \lambda_i,
\]
since $F$ is a norm. This completes the proof. $\Box$
Using Lemma 1.5.1 along with Minkowski’s successive minima theorem, we can obtain some bounds on $\mu$ in terms of the determinant of $\Lambda$ and volume of $M$. A nice bound can be easily obtained in an important special case.

**Corollary 1.5.2.** If $\lambda_1 \geq 1$, then

$$\mu \leq \frac{2^{n-1} \lambda_1 \det(\Lambda)}{\text{Vol}(M)}.$$  

**Proof.** Since

$$1 \leq \lambda_1 \leq \cdots \leq \lambda_n,$$

Theorem 1.4.2 implies

$$\lambda_n \leq \lambda_1 \cdots \lambda_n \leq \frac{2^n \det(\Lambda)}{\text{Vol}(M)},$$

and by Lemma 1.5.1,

$$\mu \leq \frac{1}{2} \sum_{i=1}^{n} \lambda_i \leq \frac{n}{2} \lambda_n.$$  

The result follows by combining these two inequalities. \qed

A general bound depending also on $\lambda_1$ was obtained by Scherk [Sch50], once again using Minkowski’s successive minima theorem (Theorem 1.4.2) and Jarnik’s inequality (Lemma 1.5.1). He observed that if $\lambda_1$ is fixed and $\lambda_2, \ldots, \lambda_n$ are subject to the conditions

$$\lambda_1 \leq \cdots \leq \lambda_n, \quad \lambda_1 \cdots \lambda_n \leq \frac{2^n \det(\Lambda)}{\text{Vol}(M)},$$

then the maximum of the sum

$$\lambda_1 + \cdots + \lambda_n$$

is attained when

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1}, \quad \lambda_n = \frac{2^n \det(\Lambda)}{\lambda_1^{n-1} \text{Vol}(M)}.$$  

Hence we obtain Scherk’s inequality for $\mu$.

**Corollary 1.5.3.**

$$\mu \leq \frac{n-1}{2} \lambda_1 + \frac{2^{n-1} \det(\Lambda)}{\lambda_1^{n-1} \text{Vol}(M)}.$$  

One can also obtain lower bounds for $\mu$. First notice that for every $\sigma \geq \mu$, then the bodies $\sigma M + x$ cover $\mathbb{R}^n$ as $x$ ranges through $\Lambda$. This means that $\mu M$ must contain a fundamental domain $F$ of $\Lambda$, and so

$$\text{Vol}(\mu M) = \mu^n \text{Vol}(M) \geq \text{Vol}(F) = \det(A),$$

hence

$$\mu \geq \left( \frac{\det(\Lambda)}{\text{Vol}(M)} \right)^{1/n}.$$  

In fact, by Theorem 1.4.2,

$$\left( \frac{\det(\Lambda)}{\text{Vol}(M)} \right)^{1/n} \geq \frac{(\lambda_1 \cdots \lambda_n)^{1/n}}{2} \geq \frac{\lambda_1}{2},$$
and combining this with (1.11), we obtain

\[(1.12) \quad \mu \geq \frac{\lambda_1}{2}.\]

Jarnik obtained a considerably better lower bound for \(\mu\) in [Jar41].

**Lemma 1.5.4.**

\[\mu \geq \frac{\lambda_n}{2}.\]

**Proof.** Let \(u_1, \ldots, u_n\) be the linearly independent points of \(\Lambda\) corresponding to the successive minima \(\lambda_1, \ldots, \lambda_n\) of \(M\) with respect to \(\Lambda\). Let \(F\) be the norm of \(M\), then

\[F(u_i) = \lambda_i, \quad \forall 1 \leq i \leq n.\]

We will first prove that for every \(x \in \Lambda\),

\[(1.13) \quad F\left(x - \frac{1}{2}u_n\right) \geq \frac{1}{2}\lambda_n.\]

Suppose not, then there exists some \(x \in \Lambda\) such that \(F\left(x - \frac{1}{2}u_n\right) < \frac{1}{2}\lambda_n\). Since \(F\) is a norm, we have

\[F(x) \leq F\left(x - \frac{1}{2}u_n\right) + F\left(\frac{1}{2}u_n\right) < \frac{1}{2}\lambda_n + \frac{1}{2}\lambda_n = \lambda_n,\]

and similarly

\[F(u_n - x) \leq F\left(\frac{1}{2}u_n - x\right) + F\left(\frac{1}{2}u_n\right) < \lambda_n.\]

Therefore, by definition of \(\lambda_n\),

\[x, u_n - x \in \text{span}_R \{u_1, \ldots, u_{n-1}\},\]

and so \(u_n = x + (u_n - x) \in \text{span}_R \{u_1, \ldots, u_{n-1}\}\), which is a contradiction. Hence we proved (1.13) for all \(x \in \Lambda\). Further, by Problem 1.16,

\[\mu = \max_{z \in \mathbb{R}^n} \min_{x \in \Lambda} F(x - z).\]

Then lemma follows by combining this observation with (1.13). \(\square\)

We define the *inhomogeneous minimum* of \(\Lambda\) to be the inhomogeneous minimum of the closed unit ball \(B_n\) with respect to \(\Lambda\), since it will occur quite often. This is another invariant of the lattice.
1.6. Problems

Problem 1.1. Let $a_1, \ldots, a_r \in \mathbb{R}^n$ be linearly independent points. Prove that $r \leq n$.

Problem 1.2. Prove that if $\Lambda$ is a lattice of rank $r$ in $\mathbb{R}^n$, $1 \leq r \leq n$, then $\text{span}_{\mathbb{R}} \Lambda$ is a subspace of $\mathbb{R}^n$ of dimension $r$ (by $\text{span}_{\mathbb{R}} \Lambda$ we mean the set of all finite real linear combinations of vectors from $\Lambda$).

Problem 1.3. Let $\Lambda$ be a lattice of rank $r$ in $\mathbb{R}^n$. By Problem 1.2, $V = \text{span}_{\mathbb{R}} \Lambda$ is an $r$-dimensional subspace of $\mathbb{R}^n$. Prove that $\Lambda$ is a discrete co-compact subset of $V$.

Problem 1.4. Let $\Lambda$ be a lattice of rank $r$ in $\mathbb{R}^n$, and let $V = \text{span}_{\mathbb{R}} \Lambda$ be an $r$-dimensional subspace of $\mathbb{R}^n$, as in Problem 1.3 above. Prove that $\Lambda$ and $V$ are both additive groups, and $\Lambda$ is a subgroup of $V$.

Problem 1.5. Let $\Lambda$ be a lattice and $\Omega$ a subset of $\Lambda$. Prove that $\Omega$ is a sublattice of $\Lambda$ if and only if it is a subgroup of the abelian group $\Lambda$.

Problem 1.6. Let $\Lambda$ be a lattice and $\Omega$ a sublattice of $\Lambda$ of the same rank. Prove that two cosets $x + \Omega$ and $y + \Omega$ of $\Omega$ in $\Lambda$ are equal if and only if $x - y \in \Omega$. Conclude that a coset $x + \Omega$ is equal to $\Omega$ if and only if $x \in \Omega$.

Problem 1.7. Let $\Lambda$ be a lattice and $\Omega \subseteq \Lambda$ a sublattice. Suppose that the quotient group $\Lambda/\Omega$ is finite. Prove that rank of $\Omega$ is the same as rank of $\Lambda$.

Problem 1.8. Given a lattice $\Lambda$ and a real number $\mu$, define $\mu \Lambda = \{\mu x : x \in \Lambda\}$. Prove that $\mu \Lambda$ is a lattice. Prove that if $\mu$ is an integer, then $\mu \Lambda$ is a sublattice of $\Lambda$.

Problem 1.9. Prove that it is possible to select the coefficients $v_{ij}$ in Theorem 1.2.5 so that the matrix $(v_{ij})_{1 \leq i, j \leq n}$ is upper (or lower) triangular with non-negative entries, and the largest entry of each row (or column) is on the diagonal.

Problem 1.10. Prove that for every point $x \in \mathbb{R}^n$ there exists uniquely a point $y \in F$ such that $x - y \in \Lambda$, i.e. $x$ lies in the coset $y + \Lambda$ of $\Lambda$ in $\mathbb{R}^n$. This means that $F$ is a full set of coset representatives of $\Lambda$ in $\mathbb{R}^n$.

Problem 1.11. Prove that volume of a fundamental parallelootope is equal to the determinant of the lattice.
Problem 1.12. Let $S$ be a compact convex set in $\mathbb{R}^n$, $A \in \text{GL}_n(\mathbb{R})$, and define

$$ T = AS = \{ Ax : x \in S \}. $$

Prove that $\text{Vol}(T) = |\det(A)| \text{Vol}(S)$. 

Hint: If we treat multiplication by $A$ as coordinate transformation, prove that its Jacobian is equal to $\det(A)$. Now use it in the integral for the volume of $T$ to relate it to the volume of $S$.

Problem 1.13. Prove versions of Theorems 1.3.1 - 1.3.2 where $\mathbb{Z}^n$ is replaced by an arbitrary lattice $\Lambda \subseteq \mathbb{R}^n$ or rank $n$ and the lower bounds on volume of $M$ are multiplied by $\det(\Lambda)$. 

Hint: Let $\Lambda = AZ^n$ for some $A \in \text{GL}_n(\mathbb{R})$. Then a point $x \in A^{-1}M \cap \mathbb{Z}^n$ if and only if $Ax \in M \cap \Lambda$. Now use Problem 1.12 to relate the volume of $A^{-1}M$ to the volume of $M$.

Problem 1.14. Prove that a standard simplex in $\mathbb{R}^n$ has volume $1/n!$.

Problem 1.15. Let $M \subset \mathbb{R}^n$ be a compact convex 0-symmetric set. Define a function $F : \mathbb{R}^n \to \mathbb{R}$, given by

$$ F(x) = \inf\{ a \in \mathbb{R}_{>0} : x \in aM \}, $$

for each $x \in \mathbb{R}^n$. Prove that this is a norm, i.e. it satisfies the three conditions:

1. $F(x) = 0$ if and only if $x = 0$,
2. $F(ax) = |a|F(x)$ for every $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$,
3. $F(x + y) \leq F(x) + F(y)$ for all $x, y \in \mathbb{R}^n$.

Problem 1.16. Let $F$ be a norm like in Problem 1.15. Prove that the inhomogeneous minimum of the corresponding set $M$ with respect to the full-rank lattice $\Lambda \subset \mathbb{R}^n$ satisfies

$$ \mu = \max_{z \in \mathbb{R}^n} \min_{x \in \Lambda} F(x - z). $$
CHAPTER 2

Discrete Optimization Problems

2.1. Sphere packing, covering and kissing number problems

Lattices play an important role in discrete optimization from classical problems to the modern day applications, such as theoretical computer science, digital communications, coding theory and cryptography, to name a few. We start with an overview of three old and celebrated problems that are closely related to the techniques in the geometry of numbers that we have so far developed, namely sphere packing, sphere covering and kissing number problems. An excellent comprehensive, although slightly outdated, reference on this subject is the well-known book by Conway and Sloane [CS88].

Let $n \geq 2$. Throughout this section by a sphere in $\mathbb{R}^n$ we will always mean a closed ball whose boundary is this sphere. We will say that a collection of spheres $\{B_i\}$ of radius $r$ is packed in $\mathbb{R}^n$ if

$$\text{int}(B_i) \cap \text{int}(B_j) = \emptyset, \; \forall \; i \neq j,$$

and there exist indices $i \neq j$ such that

$$\text{int}(B'_i) \cap \text{int}(B'_j) \neq \emptyset,$$

whenever $B'_i$ and $B'_j$ are spheres of radius larger than $r$ such that $B_i \subset B'_i$, $B_j \subset B'_j$. The sphere packing problem in dimension $n$ is to find how densely identical spheres can be packed in $\mathbb{R}^n$. Loosely speaking, the density of a packing is the proportion of the space occupied by the spheres. It is easy to see that the problem really reduces to finding the strategy of positioning centers of the spheres in a way that maximizes density. One possibility is to position sphere centers at the points of some lattice $\Lambda$ of full rank in $\mathbb{R}^n$; such packings are called lattice packings. Although clearly most packings are not lattices, it is not unreasonable to expect that best results may come from lattice packings; we will mostly be concerned with them.

**Definition 2.1.1.** Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank. The density of corresponding sphere packing is defined to be

$$\Delta = \Delta(\Lambda) := \frac{\text{proportion of the space occupied by spheres}}{\text{volume of one sphere}} = \frac{\text{volume of a fundamental domain of } \Lambda}{\text{volume of a fundamental domain of } \Lambda} = \frac{r^n \omega_n}{\det(\Lambda)}.$$
where $r$ is the packing radius, i.e. radius of each sphere in this lattice packing, and $\omega_n$ is the volume of a unit ball in $\mathbb{R}^n$, given by

\[
\omega_n = \begin{cases} 
\frac{\pi^k}{2^k k!} & \text{if } n = 2k \text{ for some } k \in \mathbb{Z} \\
\frac{2k+1}{2(2k+1)!} \pi^k (2k+1)! & \text{if } n = 2k + 1 \text{ for some } k \in \mathbb{Z}.
\end{cases}
\]

Hence the volume of a ball of radius $r$ in $\mathbb{R}^n$ is $\omega_n r^n$. It is easy to see that the packing radius $r$ is precisely the radius of the largest ball inscribed into the Voronoi cell $V$ of $\Lambda$, i.e. the inradius of $V$. Clearly $\Delta \leq 1$.

The first observation we can make is that the packing radius $r$ must depend on the lattice. In fact, it is easy to see that $r$ is precisely one half of the length of the shortest non-zero vector in $\Lambda$, in other words $r = \frac{\lambda_1}{2}$, where $\lambda_1$ is the first successive minimum of $\Lambda$. Therefore

\[
\Delta = \frac{\lambda_1 \omega_n}{2^n \det(\Lambda)}.
\]

It is not known whether the packings of largest density in each dimension are necessarily lattice packings, however we do have the following celebrated result of Minkowski (1905) generalized by Hlawka in (1944), which is usually known as Minkowski-Hlawka theorem.

**Theorem 2.1.1.** In each dimension $n$ there exist lattice packings with density

\[
\Delta \geq \frac{\zeta(n)}{2^{n-1}},
\]

where $\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$ is the Riemann zeta-function.

All known proofs of Theorem 2.1.1 are nonconstructive, so it is not generally known how to construct lattice packings with density as good as (2.2); in particular, in dimensions above 1000 the lattices whose existence is guaranteed by Theorem 2.1.1 are denser than all the presently known ones. We refer to [GL87] and [Cas59] for many further details on this famous theorem. Here we present a very brief outline of its proof, following [Cas53]. The first observation is that this theorem readily follows from the following result.

**Theorem 2.1.2.** Let $M$ be a convex bounded $0$-symmetric set in $\mathbb{R}^n$ with volume $< 2\zeta(n)$. Then there exists a lattice $\Lambda$ in $\mathbb{R}^n$ of determinant $1$ such that $M$ contains no points of $\Lambda$ except for $0$.

Now, to prove Theorem 2.1.2, we can argue as follows. Let $\chi_M$ be the characteristic function of the set $M$, i.e.

\[
\chi_M(x) = \begin{cases} 
1 & \text{if } x \in M \\
0 & \text{if } x \not\in M
\end{cases}
\]

for every $x \in \mathbb{R}^n$. For parameters $T$, $\xi_1, \ldots, \xi_{n-1}$ to be specified, let us define a lattice $\Lambda = \Lambda_T(\xi_1, \ldots, x_{n-1}) := \left\{ \left( T(a_1 + \xi_1 b), \ldots, T(a_{n-1} + \xi_{n-1} b), T^{-(n-1)} b \right) : a_1, \ldots, a_{n-1}, b \in \mathbb{Z} \right\}$.
in other words

\[
\Lambda = \begin{pmatrix}
T & 0 & \ldots & 0 & \xi_1 \\
0 & T & \ldots & 0 & \xi_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & T & \xi_{n-1} \\
0 & 0 & \ldots & 0 & T^{-(n-1)}
\end{pmatrix} \mathbb{Z}^n.
\]

Hence determinant of this lattice is 1 independent of the values of the parameters. Points of \( \Lambda \) with \( b = 0 \) are of the form \((Ta_1, \ldots, Ta_{n-1}, 0)\), and so taking \( T \) to be sufficiently large we can ensure that none of them are in \( M \), since \( M \) is bounded. Thus assume that \( T \) is large enough so that the only points of \( \Lambda \) in \( M \) have \( b \neq 0 \). Notice that \( M \) contains a nonzero point of \( \Lambda \) if and only if it contains a primitive point of \( \Lambda \), where we say that \( x \in \Lambda \) is primitive if it is not a scalar multiple of another point in \( \Lambda \). The number of symmetric pairs of primitive points of \( \Lambda \) in \( M \) is given by the counting function

\[
\eta_T(\xi_1, \ldots, \xi_{n-1}) = \sum_{b > 0} \sum_{\text{gcd}(a_1, \ldots, a_{n-1}, b)=1} \chi_M \left( T(a_1 + \xi_1 b), \ldots, T(a_{n-1} + \xi_{n-1} b), T^{-(n-1)} b \right).
\]

The argument of [Cas53] then proceeds to integrate this expression over all \( 0 \leq \xi_i \leq 1 \), \( 1 \leq i \leq n-1 \), obtaining an expression in terms of the volume of \( M \). Taking a limit as \( T \to \infty \), it is then concluded that since this volume is \(< 2\zeta(n)\), the average of the counting function \( \eta_T(\xi_1, \ldots, \xi_{n-1}) \) is less than 1. Hence there must exist some lattice of the form (2.3) which contains no nonzero points in \( M \).

In general, it is not known whether lattice packings are the best sphere packings in each dimension. In fact, the only dimensions in which optimal packings are currently known are \( n = 2, 3, 8, 24 \). In case \( n = 2 \), Gauss has proved that the best possible lattice packing is given by the hexagonal lattice

\[
\Lambda_h := \begin{pmatrix}
1 & 1 \\
0 & \sqrt{3}/2
\end{pmatrix} \mathbb{Z}^2,
\]

and in 1940 L. Fejes Tóth proved that this indeed is the optimal packing (a previous proof by Axel Thue. Its density is \( \frac{\pi}{2\sqrt{3}} \approx 0.9068996821 \).

In case \( n = 3 \), it was conjectured by Kepler that the optimal packing is given by the face-centered cubic lattice

\[
\begin{pmatrix}
-1 & -1 & 0 \\
1 & -1 & 0 \\
0 & 1 & -1
\end{pmatrix} \mathbb{Z}^3.
\]

The density of this packing is \( \approx 0.74048 \). Once again, it has been shown by Gauss in 1831 that this is the densest lattice packing, however until recently it was still not proved that this is the optimal packing. The famous Kepler’s conjecture has been settled by Thomas Hales in 1998. Theoretical part of this proof is published only in 2005 [Hal05], and the lengthy computational part was published in a series of papers in the journal of Discrete and Computational Geometry (vol. 36, no. 1 (2006)).
Dimensions $n = 8$ and $n = 24$ were settled in 2016, a week apart from each other. Maryna Viazovska [Via17], building on previous work of Cohn and Elkies [CE03], discovered a “magic” function that implied optimality of the exceptional root lattice $E_8$ for packing density in $\mathbb{R}^8$. Working jointly with Cohn, Kumar, Miller and Radchenko [CKM+17], she then immediately extended her method to dimension 24, where the optimal packing density is given by the famous Leech lattice. Detailed constructions of these remarkable lattices can be found in Conway and Sloane’s book [CS88]. This outlines the currently known results for optimal sphere packing configurations in general. On the other hand, best lattice packings are known in dimensions $n \leq 8$, as well as $n = 24$. There are dimensions in which the best known packings are not lattice packings, for instance $n = 11$.

Next we give a very brief introduction to sphere covering. The problem of sphere covering is to cover $\mathbb{R}^n$ with spheres such that these spheres have the least possible overlap, i.e. the covering has smallest possible thickness. Once again, we will be most interested in lattice coverings, that is in coverings for which the centers of spheres are positioned at the points of some lattice.

**Definition 2.1.2.** Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank. The thickness $\Theta$ of corresponding sphere covering is defined to be

$$\Theta(\Lambda) = \frac{\text{average number of spheres containing a point of the space}}{\text{volume of one sphere}} = \frac{\text{volume of a fundamental domain of } \Lambda}{\frac{R^n \omega_n}{\det(\Lambda)}}$$

where $\omega_n$ is the volume of a unit ball in $\mathbb{R}^n$, given by (2.1), and $R$ is the covering radius, i.e. radius of each sphere in this lattice covering. It is easy to see that $R$ is precisely the radius of the smallest ball circumscribed around the Voronoi cell $V$ of $\Lambda$, i.e. the circumradius of $V$. Clearly $\Theta \geq 1$.

Notice that the covering radius $R$ is precisely $\mu$, the inhomogeneous minimum of the lattice $\Lambda$. Hence combining Lemmas 1.5.1 and 1.5.4 we obtain the following bounds on the covering radius in terms of successive minima of $\Lambda$:

$$\frac{\lambda_n}{2} \leq \mu = R \leq \frac{1}{2} \sum_{i=1}^{n} \lambda_i \leq \frac{n \lambda_n}{2}.$$  

The optimal sphere covering is only known in dimension $n = 2$, in which case it is given by the same hexagonal lattice (2.4), and is equal to $\approx 1.209199$. Best possible lattice coverings are currently known only in dimensions $n \leq 5$, and it is not known in general whether optimal coverings in each dimension are necessarily given by lattices. Once again, there are dimensions in which the best known coverings are not lattice coverings.

In summary, notice that both, packing and covering properties of a lattice $\Lambda$ are very much dependent on its Voronoi cell $V$. Moreover, to simultaneously optimize packing and covering properties of $\Lambda$ we want to ensure that the inradius $r$ of $V$ is largest possible and circumradius $R$ is smallest possible. This means that we want
to take lattices with the “roundest” possible Voronoi cell. This property can be expressed in terms of the successive minima of $\Lambda$: we want

$$\lambda_1 = \cdots = \lambda_n.$$ 

Lattices with these property are called well-rounded lattices, abbreviated WR; another term ESM lattices (equal successive minima) is also sometimes used. Notice that if $\Lambda$ is WR, then by Lemma 1.5.4 we have

$$r = \frac{\lambda_1}{2} = \frac{\lambda_n}{2} \leq R,$$

although it is clearly impossible for equality to hold in this inequality. Sphere packing and covering results have numerous engineering applications, among which there are applications to coding theory, telecommunications, and image processing. WR lattices play an especially important role in these fields of study.

Another closely related classical question is known as the kissing number problem: given a sphere in $\mathbb{R}^n$ how many other non-overlapping spheres of the same radius can touch it? In other words, if we take the ball centered at the origin in a sphere packing, how many other balls are adjacent to it? Unlike the packing and covering problems, the answer here is easy to obtain in dimension 2: it is 6, and we leave it as an exercise for the reader (Problem 2.1). Although the term “kissing number” is contemporary (with an allusion to billiards, where the balls are said to kiss when they bounce), the 3-dimensional version of this problem was the subject of a famous dispute between Isaac Newton and David Gregory in 1694. It was known at that time how to place 12 unit balls around a central unit ball, however the gaps between the neighboring balls in this arrangement were large enough for Gregory to conjecture that perhaps a 13-th ball can somehow be fit in. Newton thought that it was not possible. The problem was finally solved by Schütte and van der Waerden in 1953 [SvdW53] (see also [Lee56] by J. Leech, 1956), confirming that the kissing number in $\mathbb{R}^3$ is equal to 12. The only other dimensions where the maximal kissing number is known are $n = 4, 8, 24$. More specifically, if we write $\tau(n)$ for the maximal possible kissing number in dimension $n$, then it is known that

$$\tau(2) = 6, \quad \tau(3) = 12, \quad \tau(4) = 24, \quad \tau(8) = 240, \quad \tau(24) = 196560.$$ 

In many other dimensions there are good upper and lower bounds available, and the general bounds of the form

$$2^{0.076n(1+o(1))} \leq \tau(n) \leq 2^{0.401n(1+o(1))}$$

are due to Wyner, Kabatianski and Levenshtein; see [CS88] for detailed references and many further details.

A more specialized question is concerned with the maximal possible kissing number of lattices in a given dimension, i.e. we consider just the lattice packings instead of general sphere packing configurations. Here the optimal results are known in all dimensions $n \leq 8$ and dimension 24: all of the optimal lattices here are also known to be optimal for lattice packing. Further, in all dimensions where the overall maximal kissing numbers are known, they are achieved by lattices.

Let $\Lambda \subset \mathbb{R}^n$ be a lattice, then its minimal norm $|\Lambda|$ is simply its first successive minimum, i.e.

$$|\Lambda| = \min \{ \|x\| : x \in \Lambda \setminus \{0\} \}.$$
The set of minimal vectors of $\Lambda$ is then defined as
\[
S(\Lambda) = \{ x \in \Lambda : \|x\| = |\Lambda| \}.
\]
These minimal vectors are the centers of spheres of radius $|\Lambda|/2$ in the sphere packing associated to $\Lambda$ which touch the ball centered at the origin. Hence the number of these vectors, $|S(\Lambda)|$ is precisely the kissing number of $\Lambda$. One immediate observation then is that to maximize the kissing number, same as to maximize the packing density, we want to focus our attention on WR lattices: they will have at least $2n$ minimal vectors.

A matrix $U \in \text{GL}_n(\mathbb{R})$ is called orthogonal if $U^{-1} = U^\top$, and the subset of all such matrices in $\text{GL}_n(\mathbb{R})$ is
\[
O_n(\mathbb{R}) = \{ U \in \text{GL}_n(\mathbb{R}) : U^{-1} = U^\top \}.
\]
This is a subgroup of $\text{GL}_n(\mathbb{R})$ (Problem 2.4). Discrete optimization problems on the space of lattices in a given dimension, as those discussed above, are usually considered up to the equivalence relation of similarity: two lattices $L$ and $M$ of full rank in $\mathbb{R}^n$ are called similar, denoted $L \sim M$, if there exists $\alpha \in \mathbb{R}$ and an orthogonal matrix $U \in O_n(\mathbb{R})$ such that $L = \alpha U M$. This is an equivalence relation on the space of all full-rank lattices in $\mathbb{R}^n$ (Problem 2.2), and we refer to the equivalence classes under this relation as similarity classes. If lattices $L$ and $M$ are similar, then they have the same packing density, covering thickness, and kissing number (Problem 2.3). We use the perspective of similarity classes in the next section when considering lattice packing density in the plane.
2.2. Lattice packings in dimension 2

Our goal here is to prove that the best lattice packing in \( \mathbb{R}^2 \) is achieved by the hexagonal lattice \( \Lambda_h \) as defined in (2.4) above (see Figure 1). Specifically, we will prove the following theorem.

**Theorem 2.2.1.** Let \( L \) be a lattice of rank 2 in \( \mathbb{R}^2 \). Then
\[
\Delta(L) \leq \Delta(\Lambda_h) = \frac{\pi}{2\sqrt{3}} = 0.906899\ldots,
\]
and the equality holds if and only if \( L \sim \Lambda_h \).

This result was first obtained by Lagrange in 1773, however we provide a more contemporary proof here following [Fuk11]. Our strategy is to show that the problem of finding the lattice with the highest packing density in the plane can be restricted to the well-rounded lattices without any loss of generality, where the problem becomes very simple. We start by proving that vectors corresponding to successive minima in a lattice in \( \mathbb{R}^2 \) form a basis.

**Lemma 2.2.2.** Let \( \Lambda \) be a lattice in \( \mathbb{R}^2 \) with successive minima \( \lambda_1 \leq \lambda_2 \) and let \( x_1, x_2 \) be the vectors in \( \Lambda \) corresponding to \( \lambda_1, \lambda_2 \), respectively. Then \( x_1, x_2 \) form a basis for \( \Lambda \).

**Proof.** Let \( y_1 \in \Lambda \) be a shortest vector extendable to a basis in \( \Lambda \), and let \( y_2 \in \Lambda \) be a shortest vector such that \( y_1, y_2 \) is a basis of \( \Lambda \). By picking \( \pm y_1, \pm y_2 \) if necessary we can ensure that the angle between these vectors is no greater than \( \pi/2 \). Then
\[
0 < \|y_1\| \leq \|y_2\|,
\]
and for any vector \( z \in \Lambda \) with \( \|z\| < \|y_2\| \) the pair \( y_1, z \) is not a basis for \( \Lambda \). Since \( x_1, x_2 \in \Lambda \), there must exist integers \( a_1, a_2, b_1, b_2 \) such that
\[
(x_1, x_2) = (y_1, y_2) \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}.
\]
Let \( \theta_x \) be the angle between \( x_1, x_2 \), and \( \theta_y \) be the angle between \( y_1, y_2 \), then \( \pi/3 \leq \theta_x \leq \pi/2 \) by Problem 2.6. Moreover, \( \pi/3 \leq \theta_y \leq \pi/2 \): indeed, suppose
θ_y < \pi/3, then by Problem 2.5,
\|y_1 - y_2\| < \|y_2\|,
however \(y_1, y_1 - y_2\) is a basis for \(\Lambda\) since \(y_1, y_2\) is; this contradicts the choice of \(y_2\).
Define
\[ D = \left| \det \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} \right|, \]
then \(D\) is a positive integer, and taking determinants of both sides of (2.5), we obtain
\[ (2.6) \quad \|x_1\| \|x_2\| \sin \theta_x = D \|y_1\| \|y_2\| \sin \theta_y. \]
Notice that by definition of successive minima, \(\|x_1\| \|x_2\| \leq \|y_1\| \|y_2\|\), and hence (2.6) implies that
\[ D = \frac{\|x_1\| \|x_2\| \sin \theta_x}{\|y_1\| \|y_2\| \sin \theta_y} \leq \frac{2}{\sqrt{3}} < 2, \]
meaning that \(D = 1\). Combining this observation with (2.5), we see that
\[ (x_1 \ x_2) \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}^{-1} = (y_1 \ y_2), \]
where the matrix \(\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}^{-1}\) has integer entries. Therefore \(x_1, x_2\) is also a basis for \(\Lambda\), completing the proof. □

As we know from Remark 1.4.1 in Section 1.4, the statement of Lemma 2.2.2 does not generally hold for \(d \geq 5\). We will call a basis for a lattice as in Lemma 2.2.2 a minimal basis. The goal of the next three lemmas is to show that the lattice packing density function \(\Delta\) attains its maximum in \(\mathbb{R}^2\) on the set of well-rounded lattices.

**Lemma 2.2.3.** Let \(\Lambda\) and \(\Omega\) be lattices of full rank in \(\mathbb{R}^2\) with successive minima \(\lambda_1(\Lambda), \lambda_2(\Lambda)\) and \(\lambda_1(\Omega), \lambda_2(\Omega)\) respectively. Let \(x_1, x_2\) and \(y_1, y_2\) be vectors in \(\Lambda\) and \(\Omega\), respectively, corresponding to successive minima. Suppose that \(x_1 = y_1\), and angles between the vectors \(x_1, x_2\) and \(y_1, y_2\) are equal, call this common value \(\theta\). Suppose also that
\[ \lambda_1(\Lambda) = \lambda_2(\Lambda). \]
Then
\[ \Delta(\Lambda) \geq \Delta(\Omega). \]

**Proof.** By Lemma 2.2.2, \(x_1, x_2\) and \(y_1, y_2\) are minimal bases for \(\Lambda\) and \(\Omega\), respectively. Notice that
\[ \lambda_1(\Lambda) = \lambda_2(\Lambda) = \|x_1\| = \|x_2\| = \|y_1\| = \lambda_1(\Omega) \leq \|y_2\| = \lambda_2(\Omega). \]
Then
\[ \Delta(\Lambda) = \frac{\pi \lambda_1(\Lambda)^2}{4 \det(\Lambda)} = \frac{\lambda_1(\Lambda)^2 \pi}{4 \|x_1\| \|x_2\| \sin \theta} = \frac{\pi}{4 \sin \theta} \]
\[ \geq \frac{\lambda_1(\Omega)^2 \pi}{4 \|y_1\| \|y_2\| \sin \theta} = \frac{\lambda_1(\Omega)^2 \pi}{4 \det(\Omega)} = \Delta(\Omega). \]

□
The following lemma is a converse to Problem 2.6.

**Lemma 2.2.4.** Let \( \Lambda \subset \mathbb{R}^2 \) be a lattice of full rank, and let \( x_1, x_2 \) be a basis for \( \Lambda \) such that
\[
\|x_1\| = \|x_2\|,
\]
and the angle \( \theta \) between these vectors lies in the interval \([\pi/3, \pi/2]\). Then \( x_1, x_2 \) is a minimal basis for \( \Lambda \). In particular, this implies that \( \Lambda \) is WR.

**Proof.** Let \( z \in \Lambda \), then \( z = ax_1 + bx_2 \) for some \( a, b \in \mathbb{Z} \). Then
\[
\|z\|^2 = a^2\|x_1\|^2 + b^2\|x_2\|^2 + 2abx_1^\top x_2 = (a^2 + b^2 + 2ab \cos \theta)\|x_1\|^2.
\]
If \( ab \geq 0 \), then clearly \( \|z\|^2 \geq \|x_1\|^2 \). Now suppose \( ab < 0 \), then again
\[
\|z\|^2 \geq (a^2 + b^2 - |ab|)\|x_1\|^2 \geq \|x_1\|^2,
\]
since \( \cos \theta \leq 1/2 \). Therefore \( x_1, x_2 \) are shortest nonzero vectors in \( \Lambda \), hence they correspond to successive minima, and so form a minimal basis. Thus \( \Lambda \) is WR, and this completes the proof. \( \square \)

**Lemma 2.2.5.** Let \( \Lambda \) be a lattice in \( \mathbb{R}^2 \) with successive minima \( \lambda_1, \lambda_2 \) and corresponding basis vectors \( x_1, x_2 \), respectively. Then the lattice
\[
\Lambda_{WR} = \left( x_1 \frac{\lambda_1}{\lambda_2} x_2 \right) \mathbb{Z}^2
\]
is WR with successive minima equal to \( \lambda_1 \).

**Proof.** By Problem 2.6, the angle \( \theta \) between \( x_1 \) and \( x_2 \) is in the interval \([\pi/3, \pi/2]\), and clearly this is the same as the angle between the vectors \( x_1 \) and \( \frac{\lambda_1}{\lambda_2} x_2 \). Then by Lemma 2.2.4, \( \Lambda_{WR} \) is WR with successive minima equal to \( \lambda_1 \). \( \square \)

Now combining Lemma 2.2.3 with Lemma 2.2.5 implies that
\[
(2.8) \quad \Delta(\Lambda_{WR}) \geq \Delta(\Lambda)
\]
for any lattice \( \Lambda \subset \mathbb{R}^2 \), and (2.7) readily implies that the equality in (2.8) occurs if and only if \( \Lambda = \Lambda_{WR} \), which happens if and only if \( \Lambda \) is well-rounded. Therefore the maximum packing density among lattices in \( \mathbb{R}^2 \) must occur at a WR lattice, and so for the rest of this section we talk about WR lattices only. Next observation is that for any WR lattice \( \Lambda \) in \( \mathbb{R}^2 \), (2.7) implies:
\[
\sin \theta = \frac{\pi}{4\Delta(\Lambda)},
\]
meaning that \( \sin \theta \) is an invariant of \( \Lambda \), and does not depend on the specific choice of the minimal basis. Since by our conventional choice of the minimal basis and Problem 2.6, this angle \( \theta \) is in the interval \([\pi/3, \pi/2]\), it is also an invariant of the lattice, and we call it the angle of \( \Lambda \), denoted by \( \theta(\Lambda) \).

**Lemma 2.2.6.** Let \( \Lambda \) be a WR lattice in \( \mathbb{R}^2 \). A lattice \( \Omega \subset \mathbb{R}^2 \) is similar to \( \Lambda \) if and only if \( \Omega \) is also WR and \( \theta(\Lambda) = \theta(\Omega) \).

**Proof.** First suppose that \( \Lambda \) and \( \Omega \) are similar. Let \( x_1, x_2 \) be the minimal basis for \( \Lambda \). There exist a real constant \( \alpha \) and a real orthogonal \( 2 \times 2 \) matrix \( U \) such that \( \Omega = \alpha U \Lambda \). Let \( y_1, y_2 \) be a basis for \( \Omega \) such that
\[
(y_1 \ y_2) = \alpha U(x_1 \ x_2).
\]
Then \(\|y_1\| = \|y_2\|\), and the angle between \(y_1\) and \(y_2\) is \(\theta(\Lambda) \in [\pi/3, \pi/2]\). By Lemma 2.2.4 it follows that \(y_1, y_2\) is a minimal basis for \(\Omega\), and so \(\Omega\) is WR and \(\theta(\Omega) = \theta(\Lambda)\).

Next assume that \(\Omega\) is WR and \(\theta(\Omega) = \theta(\Lambda)\). Let \(\lambda(\Lambda)\) and \(\lambda(\Omega)\) be the respective values of successive minima of \(\Lambda\) and \(\Omega\). Let \(x_1, x_2\) and \(y_1, y_2\) be the minimal bases for \(\Lambda\) and \(\Omega\), respectively. Define

\[
z_1 = \frac{\lambda(\Lambda)}{\lambda(\Omega)} y_1, \quad z_2 = \frac{\lambda(\Lambda)}{\lambda(\Omega)} y_2.
\]

Then \(x_1, x_2\) and \(z_1, z_2\) are pairs of points on the circle of radius \(\lambda(\Lambda)\) centered at the origin in \(\mathbb{R}^2\) with equal angles between them. Therefore, there exists a \(2 \times 2\) real orthogonal matrix \(U\) such that

\[
(y_1, y_2) = \frac{\lambda(\Lambda)}{\lambda(\Omega)} (z_1, z_2) = \frac{\lambda(\Lambda)}{\lambda(\Omega)} U(x_1, x_2),
\]

and so \(\Lambda\) and \(\Omega\) are similar lattices. This completes the proof. \(\square\)

We are now ready to prove the main result of this section.

**Proof of Theorem 2.2.1.** The density inequality (2.8) says that the largest lattice packing density in \(\mathbb{R}^2\) is achieved by some WR lattice \(\Lambda\), and (2.7) implies that

\[
\Delta(\Lambda) = \frac{\pi}{4 \sin \theta(\Lambda)},
\]

meaning that a smaller \(\sin \theta(\Lambda)\) corresponds to a larger \(\Delta(\Lambda)\). Problem 2.6 implies that \(\theta(\Lambda) \geq \pi/3\), meaning that \(\sin \theta(\Lambda) \geq \sqrt{3}/2\). Notice that if \(\Lambda\) is the hexagonal lattice

\[
\Lambda_h = \left(\begin{array}{cc} 1 & \frac{1}{2} \\
0 & \frac{\sqrt{3}}{2} \end{array}\right) \mathbb{Z}^2,
\]

then \(\sin \theta(\Lambda) = \sqrt{3}/2\), meaning that the angle between the basis vectors \((1, 0)\) and \((1/2, \sqrt{3}/2)\) is \(\theta = \pi/3\), and so by Lemma 2.2.4 this is a minimal basis and \(\theta(\Lambda) = \pi/3\). Hence the largest lattice packing density in \(\mathbb{R}^2\) is achieved by the hexagonal lattice. This value now follows from (2.9).

Now suppose that for some lattice \(\Lambda\), \(\Delta(\Lambda) = \Delta(\Lambda_h)\), then by (2.8) and a short argument after it \(\Lambda\) must be WR, and so

\[
\Delta(\Lambda) = \frac{\pi}{4 \sin \theta(\Lambda)} = \Delta(\Lambda_h) = \frac{\pi}{4 \sin \pi/3}.
\]

Then \(\theta(\Lambda) = \pi/3\), and so \(\Lambda\) is similar to \(\Lambda_h\) by Lemma 2.2.6. This completes the proof. \(\square\)

While we have only settled the question of best lattice packing in dimension two, we saw that well-roundedness is an essential property for a lattice to be a good contender for optimal packing density. There are, however, infinitely many WR lattices in the plane, even up to similarity, and only one of them worked well. One can then ask what properties must a lattice have to maximize packing density?
A full-rank lattice \( \Lambda \) in \( \mathbb{R}^n \) with minimal vectors \( \mathbf{x}_1, \ldots, \mathbf{x}_m \) is called eutactic if there exist positive real numbers \( c_1, \ldots, c_m \) such that

\[
\|\mathbf{v}\|^2 = \sum_{i=1}^{m} c_i (\mathbf{v}^\top \mathbf{x}_i)^2
\]

for every vector \( \mathbf{v} \in \text{span}_R \Lambda \). If \( c_1 = \cdots = c_n \), \( \Lambda \) is called strongly eutactic. A lattice is called perfect if the set of symmetric matrices

\[
\{ \mathbf{x}_i \mathbf{x}_i^\top : 1 \leq i \leq m \}
\]

spans the real vector space of \( n \times n \) symmetric matrices. These properties are preserved on similarity classes (Problem 2.7), and up to similarity there are only finitely many perfect or eutactic lattices in every dimension. For instance, up to similarity, the hexagonal lattice is the only one in the plane that is both, perfect and eutactic (Problem 2.8).

Suppose that \( \Lambda = \mathbf{A} \mathbb{Z}^n \) is a lattice with basis matrix \( \mathbf{A} \), then, as we know, \( \mathbf{B} \) is another basis matrix for \( \Lambda \) if and only if \( \mathbf{B} = \mathbf{A} \mathbf{U} \) for some \( \mathbf{U} \in \text{GL}_n(\mathbb{Z}) \). In this way, the space of full-rank lattices in \( \mathbb{R}^n \) can be identified with the set of orbits of \( \text{GL}_n(\mathbb{R}) \) under the action by \( \text{GL}_n(\mathbb{Z}) \) by right multiplication. The packing density \( \Delta \) is a continuous function on this space, and hence we can talk about its local extremum points. A famous theorem of Georgy Voronoi (1908) states that a lattice is a local maximum of the packing density function in its dimension if and only if it is perfect and eutactic. Hence, combining Problem 2.8 with Voronoi’s theorem gives another proof of unique optimality of the hexagonal lattice for lattice packing in the plane. Further, Voronoi’s theorem suggests a way of looking for the maximizer of the lattice packing density in every dimension: identify the finite set of perfect and eutactic lattices, compute their packing density and choose the largest. Unfortunately, this approach is not very practical, since already in dimension 9 the number of perfect lattices is over 9 million.
2.3. Algorithmic problems on lattices

There is a class of algorithmic problems studied in computational number theory, discrete geometry and theoretical computer science, which are commonly referred to as the lattice problems. One of their distinguishing features is that they are provably known to be very hard to solve in the sense of computational complexity of algorithms involved. Before we discuss them, let us briefly and somewhat informally recall some basic notions of computational complexity.

A key notion in theoretical computer science is that of a Turing machine as introduced by Alan Turing in 1936. Roughly speaking, this is an abstract computational device, a good practical model of which is a modern computer. It consists of an infinite tape subdivided into cells which passes through a head. The head can do the following four elementary operations: write a symbol into one cell, read a symbol from one cell, fast forward one cell, rewind one cell. These correspond to elementary operations on a computer, which uses symbols from a binary alphabet 0, 1. The number of such elementary operations required for a given algorithm is referred to as its running time. Running time is usually measured as a function of the size of the input, that is the number of cells of the infinite tape required to store the input. If we express this size as an integer \( n \) and the running time as a function \( f(n) \), then an algorithm is said to run in polynomial time if \( f(n) \) can be bounded from above by a polynomial in \( n \). We will refer to the class of problems that can be solved in polynomial time as the P class. This is our first example of a computational complexity class.

For some problems we may not know whether it is possible to solve them in polynomial time, but given a potential answer we can verify whether it is correct or not in polynomial time. Such problems are said to lie in the NP computational complexity class, where NP stands for non-deterministic polynomial. One of the most important open problems in contemporary mathematics (and arguably the most important problem in theoretical computer science) asks whether \( P = NP \)? In other words, if an answer to a problem can be verified in polynomial time, can this problem be solved by a polynomial-time algorithm? Most frequently this question is asked about decision problem, that is problems the answer to which is YES or NO. This problem, commonly known as P vs NP, was originally posed in 1971 independently by Stephen Cook and by Leonid Levin. It is believed by most experts that \( P \neq NP \), meaning that there exist problems answer to which can be verified in polynomial time, but which cannot be solved in polynomial time.

For the purposes of thinking about the P vs NP problem, it is quite helpful to introduce the following additional notions. A problem is called NP-hard if it is “at least as hard as any problem in the NP class”, meaning that for each problem in the NP class there exists a polynomial-time algorithm using which our problem can be reduced to it. A problem is called NP-complete if it is NP-hard and is known to lie in the NP class. Now suppose that we wanted to prove that \( P = NP \). One way to do this would be to find an NP-complete problem which we can show is in the \( P \) class. Since it is NP, and is at least as hard as any NP problem, this would mean that all NP problems are in the P class, and hence the equality would be proved. Although this equality seems unlikely to be true, this argument still presents serious motivation to study NP-complete problems.
As usual, we write $\Lambda \subset \mathbb{R}^n$ for a lattice of full rank and

$$0 < \lambda_1 \leq \cdots \leq \lambda_n$$

for its successive minima. A lattice can be given in the form its basis matrix, i.e. a matrix $A \in \text{GL}_n(\mathbb{R})$ such that $\Lambda = AZ^n$. There are several questions that can be asked about this setup. We formulate them in algorithmic form.

**Shortest Vector Problem (SVP).**

*Input:* A matrix $A \in \text{GL}_n(\mathbb{R})$.

*Output:* A vector $x_1 \in \Lambda = AZ^n$ such that $\|x_1\| = \lambda_1$.

**Shortest Independent Vector Problem (SIVP).**

*Input:* A matrix $A \in \text{GL}_n(\mathbb{R})$.

*Output:* Linearly independent vectors $x_1, \ldots, x_n \in \Lambda = AZ^n$ such that $\|x_i\| = \lambda_1$ for all $1 \leq i \leq n$.

**Closest Vector Problem (CVP).**

*Input:* A matrix $A \in \text{GL}_n(\mathbb{R})$ and a vector $y \in \mathbb{R}^n$.

*Output:* A vector $x \in \Lambda = AZ^n$ such that $\|x - y\| \leq \|z - y\|$ for all $z \in \Lambda$.

**Shortest Basis Problem (SBP).**

*Input:* A matrix $A \in \text{GL}_n(\mathbb{R})$.

*Output:* A basis $b_1, \ldots, b_n$ for $\Lambda = Z^n$ such that $\min \{ \|x\| : x \in \Lambda \text{ is such that } b_1, \ldots, b_{i-1}, x \text{ is extendable to a basis} \}$ for all $1 \leq i \leq n$.

Notice that SVP is a special case of CVP where the input vector $y$ is taken to be $0$: indeed, a vector corresponding to the first successive minimum is precisely a vector that is closer to the origin than any other point of $\Lambda$. On the other hand, SIVP and SBP are different problems: as we know, lattices in dimensions 5 higher may not have a basis of vectors corresponding to successive minima.

All of these algorithmic problems are all known to be NP-complete. In fact, even the problem of determining the first successive minimum of the lattice is already NP-complete. We can also ask for $\gamma$-approximate versions of these problems for some approximation factor $\gamma$. In other words, for the same input we want to return an answer that is bigger than the optimal by a factor of no more than $\gamma$. For instance, the $\gamma$-SVP would ask for a vector $x \in \Lambda$ such that $\|x\| \leq \gamma \lambda_1$.

It is an open problem to decide whether the $\gamma$-approximate versions of these problems are in the P class for any values of $\gamma$ polynomial in the dimension $n$.

On the other hand, $\gamma$-approximate versions of these problems for $\gamma$ exponential in $n$ are known to be polynomial. The most famous such approximation algorithm is LLL, which was discovered by A. Lenstra, H. Lenstra and L. Lovasz in 1982 [LLL82]. LLL is a polynomial time reduction algorithm that, given a lattice $\Lambda$, produces a basis $b_1, \ldots, b_n$ for $\Lambda$ such that

$$\min_{1 \leq i \leq n} \|b_i\| \leq 2^{\frac{n-1}{2}} \lambda_1,$$
and
\[
\prod_{i=1}^{n} \|b_i\| \leq 2^{\frac{n(n-1)}{4}} \det(\Lambda).
\]

We can compare this to the upper bound given by Minkowski’s Successive Minima Theorem (Theorem 1.4.2):
\[
\prod_{i=1}^{n} \lambda_i \leq \frac{2^n}{\omega_n} \det(\Lambda).
\]

For instance, when \( n = 2k \) the bound (2.10) gives
\[
\prod_{i=1}^{n} \|b_i\| \leq 2^{\frac{k(k-1)}{2}} \det(\Lambda),
\]
while (2.11) gives
\[
\prod_{i=1}^{n} \lambda_i \leq \frac{4^k k!}{\pi^k} \det(\Lambda).
\]

Let us briefly describe the main idea behind LLL. The first observation is that an orthogonal basis, if one exists in a lattice, is always the shortest one. Indeed, suppose \( u_1, \ldots, u_n \) is such a basis, then for any \( a_1, \ldots, a_n \in \mathbb{Z} \),
\[
\left\| \sum_{i=1}^{n} a_i u_i \right\|^2 = \sum_{i=1}^{n} a_i^2 \|u_i\|^2,
\]
which implies that the shortest basis vectors can only be obtained by taking one of the coefficients \( a_i = \pm 1 \) and the rest 0. Of course, most lattices do not have orthogonal bases, in which case finding a short basis is much harder. Still, the basic principle of constructing a short basis is based on looking for vectors that would be “close to orthogonal”.

We observed in Section 2.2 (in particular, see Problems 2.5, 2.6, Lemma 2.2.4) that the angle between a pair of shortest vectors must be between \([\pi/3, 2\pi/3]\), i.e., these vectors are “near-orthogonal”: in fact, these vectors have to be as close to orthogonal as possible within the lattice. This is the underlying idea behind the classical Lagrange-Gauss Algorithm for finding a shortest basis for a lattice in \( \mathbb{R}^2 \). Specifically, an ordered basis \( b_1, b_2 \) for a planar lattice \( \Lambda \) consists of vectors corresponding to successive minima \( \lambda_1, \lambda_2 \) of \( \Lambda \), respectively, if and only if
\[
\mu := \frac{b_1^\top b_2}{\|b_1\|^2} \leq \frac{1}{2}.
\]

On the other hand, if \(|\mu| > 1/2\), then replacing \( b_2 \) with
\[
b_2 - |\mu| b_1,
\]
where \(|\mu|\) stands for the nearest integer to \( \mu \), produces a shorter second basis vector. We leave the proof of this as an exercise (Problem 2.9). Hence we can formulate the Gauss-Lagrange Algorithm:

\textit{Input:} \( b_1, b_2 \in \mathbb{R}^2 \) such that \( \|b_1\| \leq \|b_2\| \)
\textit{Compute:} \( \mu = \frac{b_1^\top b_2}{\|b_1\|^2} \)
\textit{Check:} if \(|\mu| \leq 1/2\), output \( b_1, b_2 \); else set \( b_2 \leftarrow b_2 - |\mu| b_1 \) and repeat the algorithm (swapping \( b_1, b_2 \), if necessary, to ensure \( \|b_1\| \leq \|b_2\| \))
This algorithm terminates in a finite number of steps (Problem 2.10).

Let us demonstrate this algorithm on an example. Suppose \( \Lambda = \text{span}_\mathbb{Z}\{b_1, b_2\} \), where
\[
b_1 = \left( \begin{array}{c} 1 \\ 5 \end{array} \right), \quad b_2 = \left( \begin{array}{c} 1 \\ 0 \end{array} \right).
\]
We notice that \( \|b_1\| > \|b_2\| \), so we swap the vectors: \( b_1 \leftrightarrow b_2 \). We then compute
\[
\mu = \frac{b_1^\top b_2}{\|b_1\|^2} = 1 > 1/2.
\]
The nearest integer to \( \mu \) is 1, so we set
\[
b_2 \leftarrow b_2 - b_1 = \left( \begin{array}{c} 0 \\ 5 \end{array} \right).
\]
We still have \( \|b_1\| < \|b_2\| \), so no need to swap the vectors. With the new basis \( b_1, b_2 \) we again compute \( \mu \), which is now equal to \( 0 < 1/2 \). Hence we found a shortest basis for \( \Lambda \):
\[
\left( \begin{array}{c} 1 \\ 0 \end{array} \right), \quad \left( \begin{array}{c} 0 \\ 5 \end{array} \right).
\]

LLL is based on a generalization of this idea. We can start with a basis \( b_1, \ldots, b_n \) for a lattice \( \Lambda \) in \( \mathbb{R}^n \) and use the Gram-Schmidt orthogonalization procedure to compute a corresponding orthogonal (but not normalized) basis \( b'_1, \ldots, b'_n \) for \( \mathbb{R}^n \). For any pair of indices \( i, j \) with \( 1 \leq j < i \leq n \), let us compute the Gram-Schmidt coefficient
\[
\mu_{ij} = \frac{b_i^\top b'_j}{\|b'_j\|^2}.
\]
If this coefficient is \( > 1/2 \) in absolute value, we swap \( b_i \leftarrow b_i - \lfloor \mu \rfloor b_j \); this ensures the length reduction, but one other condition is also needed. Formally speaking, a resulting basis \( b_1, \ldots, b_n \) is called LLL reduced if the following two conditions are satisfied:

1. For all \( 1 \leq j < i \leq n \), \( |\mu_{ij}| \leq 1/2 \)
2. For some parameter \( \delta \in [1/4, 1) \), for all \( 1 \leq k \leq n \),
\[
\delta \|b'_{k-1}\|^2 \leq \|b'_k\|^2 + \mu_{k,(k-1)}^2 \|b'_{k-1}\|^2.
\]
Traditionally, \( \delta \) is taken to be 3/4. While we will not go into further details about the LLL, some good more detailed references on this subject include the original paper \([LLL82]\), as well as more recent books \([Coh00]\), \([Bor02]\), and \([HPS08]\).
2.4. Problems

Problem 2.1. Prove that the optimal kissing number in $\mathbb{R}^2$ is equal to 6.

Problem 2.2. Prove that similarity is an equivalence relation on the set of all lattices of full rank in $\mathbb{R}^n$.

Problem 2.3. Assume two full-rank lattices $L$ and $M$ in $\mathbb{R}^n$ are similar. Prove that they have the same packing density, covering thickness and kissing number.

Problem 2.4. Prove that the set of all real orthogonal $n \times n$ matrices $O_n(\mathbb{R})$ is a subgroup of $GL_n(\mathbb{R})$.

Problem 2.5. Let $x_1$ and $x_2$ be nonzero vectors in $\mathbb{R}^2$ so that the angle $\theta$ between them satisfies $0 < \theta < \frac{\pi}{3}$. Prove that
\[ \|x_1 - x_2\| < \max\{\|x_1\|, \|x_2\|\}. \]

Problem 2.6. Let $\Lambda \subset \mathbb{R}^2$ be a lattice of full rank with successive minima $\lambda_1 \leq \lambda_2$, and let $x_1, x_2$ be the vectors in $\Lambda$ corresponding to $\lambda_1, \lambda_2$, respectively. Let $\theta \in [0, \pi/2]$ be the angle between $x_1$ and $x_2$. Prove that
\[ \pi/3 \leq \theta \leq \pi/2. \]

Problem 2.7. Let $L$ and $M$ be two similar lattices. Prove that if $L$ is eutactic (respectively, strongly eutactic, perfect), then so is $M$.

Problem 2.8. Prove that the hexagonal lattice $\Lambda_h$ is both, perfect and eutactic. Further, prove that if $L$ is a perfect lattice in $\mathbb{R}^2$, then $L \sim \Lambda_h$.

Problem 2.9. Prove that an ordered basis $b_1, b_2$ for a planar lattice $\Lambda$ consists of vectors corresponding to successive minima $\lambda_1, \lambda_2$, respectively, if and only if
\[ \mu := \frac{b_1^T b_2}{\|b_1\|^2} \leq \frac{1}{2}. \]
On the other hand, if $|\mu| > 1/2$, then replacing $b_2$ with
\[ b_2 - |\mu| b_1, \]
where $|\mu|$ stands for the nearest integer to $\mu$, produces a shorter second basis vector.

Problem 2.10. Prove that the Gauss-Lagrange Algorithm as discussed in Section 2.3 terminates in a finite number of steps.
CHAPTER 3

Quadratic forms

3.1. Introduction to quadratic forms

The theory of lattices that we introduced in the previous chapters can be viewed
from a somewhat different angle, namely from the point of view of positive definite
quadratic forms. In this chapter we study some basic properties of quadratic forms
and then emphasize the connection to lattices.

A quadratic form is a homogeneous polynomial of degree 2; unless explicitly
stated otherwise, we consider quadratic forms with real coefficients. More generally,
we can talk about a symmetric bilinear form, that is a polynomial

\[ B(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} X_i Y_j, \]

in 2n variables \(X_1, \ldots, X_n, Y_1, \ldots, Y_n\) so that \(b_{ij} = b_{ji}\) for all 1 \(\leq i, j \leq n\). Such a
polynomial \(B\) is called bilinear because although it is not linear, it is linear in each
set of variables, \(X_1, \ldots, X_n\) and \(Y_1, \ldots, Y_n\). It is easy to see that a bilinear form
\(B(X, Y)\) can also be written as

\[ B(X, Y) = X^\top B Y, \]

where

\[ B = \begin{pmatrix}
    b_{11} & b_{12} & \cdots & b_{1n} \\
    b_{12} & b_{22} & \cdots & b_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{1n} & b_{2n} & \cdots & b_{nn}
\end{pmatrix}, \]

is the corresponding \(n \times n\) symmetric coefficient matrix, called the Gram matrix
of the form, and

\[ X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \]

are the variable vectors. Hence symmetric bilinear forms are in bijective correspon-
dence with symmetric \(n \times n\) matrices. It is also easy to notice that

\[(3.1) \quad B(X, Y) = X^\top B Y = (X^\top B Y)^\top = Y^\top B^\top X = Y^\top B X = B(Y, X), \]

since \(B\) is symmetric. We can also define the corresponding quadratic form

\[ Q(X) = B(X, X) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} X_i X_j = X^\top B X. \]

Hence to each bilinear symmetric form in 2n variables there corresponds a quadratic
form in \(n\) variables. The converse is also true (Problem 3.1).
We define the determinant or discriminant of a symmetric bilinear form $B$ and of its associated quadratic form $Q$ to be the determinant of the coefficient matrix $B$, and will denote it by $\text{det}(B)$ or $\text{det}(Q)$.

Many properties of bilinear and corresponding quadratic forms can be deduced from the properties of their matrices. Hence we start by recalling some properties of symmetric matrices.

**Lemma 3.1.1.** A real symmetric matrix has all real eigenvalues.

**Proof.** Let $B$ be a real symmetric matrix, and let $\lambda$ be an eigenvalue of $B$ with a corresponding eigenvector $x$. Write $\overline{\lambda}$ for the complex conjugate of $\lambda$, and $\overline{B}$ and $\overline{x}$ for the matrix and vector correspondingly whose entries are complex conjugates of respective entries of $B$ and $x$. Then $Bx = \lambda x$, and so

$$B\overline{x} = \overline{B}x = \overline{\lambda}x = \overline{\lambda}x,$$

since $B$ is a real matrix, meaning that $B = \overline{B}$. Then, by (3.1)

$$\lambda(\overline{x}^T) = (\lambda x)^T \overline{x} = x^T \overline{B}x = x^T(\overline{\lambda}x) = \overline{\lambda}(x^T \overline{x}),$$

meaning that $\lambda = \overline{\lambda}$, since $x^T \overline{x} \neq 0$. Therefore $\lambda \in \mathbb{R}$. 

**Remark 3.1.1.** Since eigenvectors corresponding to real eigenvalues of a matrix must be real, Lemma 3.1.1 implies that a real symmetric matrix has all real eigenvalues as well. In fact, even more is true.

**Lemma 3.1.2.** Let $B$ be a real symmetric matrix. Then there exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $B$.

**Proof.** We argue by induction on $n$. If $n = 1$, the result is trivial. Hence assume $n > 1$, and the statement of the lemma is true for $n - 1$. Let $x_1$ be an eigenvector of $B$ with the corresponding eigenvalue $\lambda_1$. We can assume that $\|x_1\| = 1$. Use Gram-Schmidt orthogonalization process to extend $x_1$ to an orthonormal basis for $\mathbb{R}^n$, and write $U$ for the corresponding basis matrix such that $x_1$ is the first column. Then it is easy to notice that $U^{-1} = U^\top$. By Problem 3.2,

$$U^\top BU = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & a_{11} & \cdots & a_{1(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{(n-1)1} & \cdots & a_{(n-1)(n-1)} \end{pmatrix},$$

where the $(n-1) \times (n-1)$ matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1(n-1)} \\ \vdots & \ddots & \vdots \\ a_{(n-1)1} & \cdots & a_{(n-1)(n-1)} \end{pmatrix}$$

is also symmetric. Now we can apply induction hypothesis to the matrix $A$, thus obtaining an orthonormal basis for $\mathbb{R}^{n-1}$, consisting of eigenvectors of $A$, call them $y_2, \ldots, y_n$. For each $2 \leq i \leq n$, define

$$y_i' = \begin{pmatrix} 0 \\ y_i \end{pmatrix} \in \mathbb{R}^n.$$
and let $x_i = U y'_i$. There exist $\lambda_2, \ldots, \lambda_n$ such that $A y_i = \lambda_i y_i$ for each $2 \leq i \leq n$, hence
\[ U^\top B U y'_i = \lambda_i y'_i, \]
and so $B x_i = \lambda_i x_i$. Moreover, for each $2 \leq i \leq n$,
\[ x_i^\top x_i = (x_i^\top U) \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0, \]
by construction of $U$. Finally notice that for each $2 \leq i \leq n$,
\[ \|x_i\| = \left( U \begin{pmatrix} 0 \\ y_i \end{pmatrix} \right)^\top U \begin{pmatrix} 0 \\ y_i \end{pmatrix} = \|y_i\| = 1, \]
meaning that $x_1, x_2, \ldots, x_n$ is precisely the basis we are looking for. □

Remark 3.1.2. An immediate implication of Lemma 3.1.2 is that a real symmetric matrix has $n$ linearly independent eigenvectors, hence is diagonalizable; we will prove an even stronger statement below. In particular, this means that for each eigenvalue, its algebraic multiplicity (i.e. multiplicity as a root of the characteristic polynomial) is equal to its geometric multiplicity (i.e. dimension of the corresponding eigenspace).

Lemma 3.1.3. Every real symmetric matrix $B$ is diagonalizable by an orthogonal matrix, i.e. there exists a matrix $U \in O_n(\mathbb{R})$ such that $U^\top B U$ is a diagonal matrix.

Proof. By Lemma 3.1.2, we can pick an orthonormal basis $u_1, \ldots, u_n$ for $\mathbb{R}^n$ consisting of eigenvectors of $B$. Let
\[ U = (u_1 \ldots u_n), \]
be the corresponding orthogonal matrix. Then for each $1 \leq i \leq n$,
\[ u_i^\top B u_i = u_i^\top (\lambda_i u_i) = \lambda_i (u_i^\top u_i) = \lambda_i, \]
where $\lambda_i$ is the corresponding eigenvalue, since
\[ 1 = \|u_i\|^2 = u_i^\top u_i. \]
Also, for each $1 \leq i \neq j \leq n$,
\[ u_i^\top B u_j = u_i^\top (\lambda_j u_j) = \lambda_j (u_i^\top u_j) = 0. \]
Therefore, $U^\top B U$ is a diagonal matrix whose diagonal entries are precisely the eigenvalues of $B$. □

Remark 3.1.3. Lemma 3.1.3 is often referred to as the Principal Axis Theorem. The statements of Lemmas 3.1.1, 3.1.2, and 3.1.3 together are usually called the Spectral Theorem for symmetric matrices; it has many important applications in various areas of mathematics, especially in Functional Analysis, where it is usually interpreted as a statement about self-adjoint (or hermitian) linear operators. A more general version of Lemma 3.1.3, asserting that any matrix is unitary-similar to an upper triangular matrix over an algebraically closed field, is usually called Schur’s theorem.

We now discuss the implications of these results for quadratic forms. A linear transformation $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is a homomorphism of additive groups, which is an isomorphism if and only if its matrix is nonsingular (Problem 3.3). We will call such
homomorphisms (respectively, isomorphisms) linear. Then $\text{GL}_n(\mathbb{R})$ is precisely the group of all linear isomorphisms $\mathbb{R}^n \to \mathbb{R}^n$.

Remark 3.1.4. Interestingly, not all homomorphisms $\mathbb{R}^n \to \mathbb{R}^n$ are linear: in fact, in a certain sense most of them are not linear. Nonlinear homomorphisms from $\mathbb{R}^n$ to $\mathbb{R}^n$, however, cannot be explicitly constructed. This has to do with the fact that a basis for $\mathbb{R}$ as $\mathbb{Q}$-vector space (called the Hamel basis), while has to exist by the Axiom of Choice, cannot be explicitly constructed (see [Kuc09] for details).

Definition 3.1.2. Two real symmetric bilinear forms $B_1$ and $B_2$ in $2n$ variables are called isomorphic if there exists an isomorphism $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$B_1(\sigma x, \sigma y) = B_2(x, y),$$

for all $x, y \in \mathbb{R}^n$. Their associated quadratic forms $Q_1$ and $Q_2$ are also said to be isomorphic in this case and $\sigma$ is called an isomorphism of these bilinear (respectively, quadratic) forms.

Isomorphism is easily seen to be an equivalence relation on real symmetric bilinear (respectively quadratic) forms, so we can talk about isomorphism classes of real symmetric bilinear (respectively quadratic) forms. The set of all isomorphisms from a bilinear form $B$ to itself forms a group under matrix multiplication, which is a subgroup of $\text{GL}_n(\mathbb{R})$ (Problem 3.4): these are precisely the linear maps $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$B(\sigma X, \sigma Y) = B(X, Y),$$

and so the same is true for the associated quadratic form $Q$.

Definition 3.1.3. A symmetric bilinear or quadratic form is called nonsingular (or nondegenerate, or regular) if its Gram matrix is nonsingular. Alternative equivalent characterizations of nonsingular forms are given in Problem 3.5. We now deal with nonsingular quadratic forms until further notice.
Definition 3.1.5. A nonsingular diagonal quadratic form $Q$ can be written as

$$Q(X) = \sum_{j=1}^{r} b_{ij} X_{ij}^2 - \sum_{j=1}^{s} b_{kj} X_{kj}^2,$$

where all coefficients $b_{ij}, b_{kj}$ are positive. In other words, $r$ of the diagonal terms are positive, $s$ are negative, and $r + s = n$. The pair $(r, s)$ is called the signature of $Q$. If $Q$ is a non-diagonal nonsingular quadratic form, we define its signature to be the signature of an isometric diagonal form.

The following is Lemma 5.4.3 on p. 333 of [Jac90]; the proof is essentially the same.

Theorem 3.1.5. Signature of a nonsingular quadratic form is uniquely determined.

Proof. We will show that signature of a nonsingular quadratic form $Q$ does not depend on the choice of diagonalization. Let $B$ be the coefficient matrix of $Q$, and let $U, W$ be two different matrices that diagonalize $B$ with column vectors $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$, respectively, arranged in such a way that

$$Q(u_1), \ldots, Q(u_{r_1}) > 0, \quad Q(u_{r_1+1}), \ldots, Q(u_n) < 0,$$

and

$$Q(w_1), \ldots, Q(w_{r_2}) > 0, \quad Q(w_{r_2+1}), \ldots, Q(w_n) < 0,$$

for some $r_1, r_2 \leq n$. Define vector spaces

$$V_1^+ = \text{span}_\mathbb{R}\{u_1, \ldots, u_{r_1}\}, \quad V_1^- = \text{span}_\mathbb{R}\{u_{r_1+1}, \ldots, u_n\},$$

and

$$V_2^+ = \text{span}_\mathbb{R}\{w_1, \ldots, w_{r_2}\}, \quad V_2^- = \text{span}_\mathbb{R}\{w_{r_2+1}, \ldots, w_n\}.$$

Clearly, $Q$ is positive on $V_1^+, V_2^+$ and is negative on $V_1^-, V_2^-$. Therefore,

$$V_1^+ \cap V_2^- = V_2^+ \cap V_1^- = \{0\}.$$

Then we have

$$r_1 + (n - r_2) = \dim(V_1^+ \oplus V_2^-) \leq n,$$

and

$$r_2 + (n - r_1) = \dim(V_2^+ \oplus V_1^-) \leq n,$$

which implies that $r_1 = r_2$. This completes the proof. \qed

The importance of signature for nonsingular real quadratic forms is that it is an invariant not just of the form itself, but of its whole isometry class.

Theorem 3.1.6 (Sylvester’s Theorem). Two nonsingular real quadratic forms in $n$ variables are isomorphic if and only if they have the same signature.

We leave the proof of this theorem to exercises (Problem 3.6). An immediate implication of Theorem 3.1.6 is that for each $n \geq 2$, there are precisely $n+1$ isomorphism classes of nonsingular real quadratic forms in $n$ variables, and by Theorem 3.1.4 each of these classes contains a diagonal form. Some of these isomorphism classes are especially important for our purposes.

Definition 3.1.6. A quadratic form $Q$ is called positive or negative definite if, respectively, $Q(x) > 0$, or $Q(x) < 0$ for each $0 \neq x \in \mathbb{R}^n$; $Q$ is called positive or negative semi-definite if, respectively, $Q(x) \geq 0$, or $Q(x) \leq 0$ for each $0 \neq x \in \mathbb{R}^n$. Otherwise, $Q$ is called indefinite.
A real quadratic form is positive (respectively, negative) definite if and only if it has signature \((n,0)\) (respectively, \((0,n)\)). In particular, a definite form has to be nonsingular (Problem 3.7). Positive definite real quadratic forms are also sometimes called norm forms, since they do define norms (Problem 3.8).

We now have the necessary machinery to relate quadratic forms to lattices. Let \(\Lambda\) be a lattice of full rank in \(\mathbb{R}^n\), and let \(A\) be a basis matrix for \(\Lambda\). Then \(y \in \Lambda\) if and only if \(y = Ax\) for some \(x \in \mathbb{Z}^n\). Notice that the Euclidean norm of \(y\) in this case is \(\|y\| = (Ax)\top (Ax) = x\top (A\top A)x = Q_A(x)\), where \(Q_A\) is the quadratic form whose Gram matrix is \(A\top A\). By construction, \(Q_A\) must be a positive definite form. This quadratic form is called the norm form for the lattice \(\Lambda\) corresponding to the basis matrix \(A\).

Now suppose \(C\) is another basis matrix for \(\Lambda\). Then there must exist \(U \in \text{GL}_n(\mathbb{Z})\) such that \(C = AU\). Hence the matrix of the quadratic form \(Q_C\) is \((AU)\top (AU) = U\top (A\top A)U\); we call two such matrices \(\text{GL}_n(\mathbb{Z})\)-congruent. Notice in this case that for each \(x \in \mathbb{R}^n\)

\[
Q_C(x) = x\top U\top (A\top A)Ux = Q_A(Ux),
\]

which means that the quadratic forms \(Q_A\) and \(Q_C\) are isomorphic. In such cases, when there exists an isomorphism between two quadratic forms in \(\text{GL}_n(\mathbb{Z})\), we will call them arithmetically equivalent. We proved the following statement.

**Proposition 3.1.7.** All different norm forms of a lattice \(\Lambda\) of full rank in \(\mathbb{R}^n\) are arithmetically equivalent to each other.

Moreover, suppose that \(Q\) is a positive definite quadratic form with Gram matrix \(B\), then there exists \(U \in \mathcal{O}_n(\mathbb{R})\) such that

\[
U\top BU = D,
\]

where \(D\) is a nonsingular diagonal \(n \times n\) matrix with positive entries on the diagonal. Write \(\sqrt{D}\) for the diagonal matrix whose entries are positive square roots of the entries of \(D\), then \(D = \sqrt{D} \cdot \sqrt{D}\), and so

\[
\mathcal{B} = (\sqrt{D}U)\top (\sqrt{D}U).
\]

Letting \(A = \sqrt{D}U\) and \(\Lambda = AZ^n\), we see that \(Q\) is a norm form of \(\Lambda\). Notice that the matrix \(A\) is unique only up to orthogonal transformations, i.e. for any \(W \in \mathcal{O}_n(\mathbb{R})\)

\[
(WA)\top (WA) = A\top (W\top W)A = A\top A = B.
\]

Therefore \(Q\) is a norm form for every lattice \(WAZ^n\), where \(W \in \mathcal{O}_n(\mathbb{R})\). Let us call two lattices \(\Lambda_1\) and \(\Lambda_2\) isometric if there exists \(W \in \mathcal{O}_n(\mathbb{R})\) such that \(\Lambda_1 = W\Lambda_2\). This is easily seen to be an equivalence relation on lattices. Hence we have proved the following.

**Theorem 3.1.8.** Arithmetic equivalence classes of real positive definite quadratic forms in \(n\) variables are in bijective correspondence with isometry classes of full rank lattices in \(\mathbb{R}^n\).

Notice in particular that if a lattice \(\Lambda\) and a quadratic form \(Q\) correspond to each other as described in Theorem 3.1.8, then

\[
\det(\Lambda) = \sqrt{|\det(Q)|}.
\]
Now that we have the bijective correspondence between lattices and positive
definite quadratic forms, we end this section with an application of Minkowski’s
Convex Body Theorem to the context of quadratic forms: this is Theorem 4 on p.
44 of [GL87].

**Theorem 3.1.9.** Let
\[
Q(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} X_i X_j = X^\top B X
\]
be a positive definite quadratic form in \(n\) variables with Gram matrix \(B\). There
exists \(0 \neq x \in \mathbb{Z}^n\) such that
\[
Q(x) \leq 4 \left( \frac{\det(B)}{\omega_n^2} \right)^{1/n}.
\]

**Proof.** As in the proof of Theorem 3.1.8 above, we can decompose \(B\) as
\(B = A^\top A\) for some \(A \in \text{GL}_n(\mathbb{R})\). Then
\[
\det(B) = \det(A)^2.
\]
For each \(r \in \mathbb{R}_{>0}\), define the set
\[
E_r = \{x \in \mathbb{R}^n : Q(x) \leq r\} = \{x \in \mathbb{R}^n : (Ax)^\top (Ax) \leq r\} = A^{-1} S_r,
\]
where \(S_r = \{y \in \mathbb{R}^n : \|y\|^2 \leq r\}\) is a ball of radius \(\sqrt{r}\) centered at the origin in
\(\mathbb{R}^n\). Hence \(E_r\) is an ellipsoid centered at the origin with
\[
\text{Vol}(E_r) = |\det(A)|^{-1} \text{Vol}(S_r) = \omega_n \sqrt{\frac{r^n}{\det(B)}}.
\]
Hence if
\[
r = 4 \left( \frac{\det(B)}{\omega_n^2} \right)^{1/n},
\]
then \(\text{Vol}(E_r) = 2^n\), and so by Theorem 1.3.2 there exists \(0 \neq x \in E_r \cap \mathbb{Z}^n\).
\(\square\)
3.2. Minkowski’s reduction

Let $M \subseteq \mathbb{R}^n$ be a $0$-symmetric convex set with positive volume, and let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank, as before. In Section 1.3 we discussed the following question: by how much should $M$ be homogeneously expanded so that it contains $n$ linearly independent points of $\Lambda$? We learned however that the resulting set of $n$ minimal linearly independent vectors produced this way is not necessarily a basis for $\Lambda$. In this section we want to understand by how much should $M$ be homogeneously expanded so that it contains a basis of $\Lambda$? We start with some definitions. In case $M$ is a unit ball, this question is directly related to the Shortest Basis Problem (SBP) we discussed in Section 2.3. There we reviewed the polynomial-time approximation algorithm LLL for SBP. Here we will discuss the Minkowski reduction, which (in case $M$ is a unit ball) yields precisely the shortest vector. Minkowski reduction, however, cannot be implemented as a polynomial-time algorithm.

As before, let us write $F$ for the norm corresponding to $M$, i.e.

$$M = \{x \in \mathbb{R}^n : F(x) \leq 1\},$$

then

$$F(x + y) \leq F(x) + F(y).$$

We write $\lambda_1, \ldots, \lambda_n$ for the successive minima of $M$ with respect to $\Lambda$.

**Definition 3.2.1.** A basis $\{v_1, \ldots, v_n\}$ of $\Lambda$ is said to be **Minkowski reduced with respect to** $M$ if for each $1 \leq i \leq n$, $v_i$ is such that

$$F(v_i) = \min \{F(v) : v_1, \ldots, v_{i-1}, v \text{ is extendable to a basis of } \Lambda\}.$$

In the frequently occurring case when $M$ is the closed unit ball $B_n$ centered at $0$, we will just say that a corresponding such basis is **Minkowski reduced**. Notice in particular that a Minkowski reduced basis contains a shortest non-zero vector in $\Lambda$.

From here on let $\{v_1, \ldots, v_n\}$ be a Minkowski reduced basis of $\Lambda$ with respect to $M$. Then

$$F(v_1) = \lambda_1, \quad F(v_i) \geq \lambda_i \quad \forall \ 2 \leq i \leq n.$$ 

Assume first that $M = B_n$, then $F = \|\|$. Write $A$ for the corresponding basis matrix of $\Lambda$, i.e. $A = (v_1 \ldots v_n)$, and so $\Lambda = AZ^n$. Let $Q$ be the corresponding positive definite quadratic form, i.e. for each $x \in \mathbb{R}^n$

$$Q(x) = x^\top A^\top Ax.$$ 

Then, as we noted before, $Q(x) = \|Ax\|^2$. In particular, for each $1 \leq i \leq n$

$$Q(e_i) = \|v_i\|^2.$$ 

Hence for each $1 \leq i \leq n$, $Q(e_i) \leq Q(x)$ for all $x$ such that

$$v_1, \ldots, v_{i-1}, Ax$$

is extendable to a basis of $\Lambda$. This means that for every $1 \leq i \leq n$

(3.3) $$Q(e_i) \leq Q(x) \quad \forall \ x \in \mathbb{Z}^n, \ \gcd(x_1, \ldots, x_n) = 1.$$ 

If a positive definite quadratic form satisfies (3.3), we will say that it is **Minkowski reduced**. Every positive definite quadratic form is arithmetically equivalent to a Minkowski reduced form (Problem 3.9).

Now let us drop the assumption that $M = B_n$, but preserve the rest of notation as above. We can prove the following analogue of Minkowski’s successive
Then and a real number $\beta$ where $\alpha$ is essentially Theorem 2 on p. 66 of [GL87], which is due to Minkowski, Mahler, and Weyl.

**Theorem 3.2.1.** Let $\nu_1 = 1$, and $\nu_i = \left(\frac{3}{2}\right)^{i-2}$ for each $2 \leq i \leq n$. Then

$$\lambda_i \leq F(v_i) \leq \nu_i \lambda_i.$$  

Moreover,

$$\prod_{i=1}^{n} F(v_i) \leq 2^n \left(\frac{3}{2}\right)^{\frac{(n-1)(n-2)}{2}} \frac{\det(\Lambda)}{\text{Vol}(M)}.$$  

**Proof.** It is easy to see that (3.5) follows immediately by combining (3.4) with Theorem 1.4.2, hence we only need to prove (3.4). It is obvious by definition of a reduced basis that $F(v_i) \geq \lambda_i$ for each $1 \leq i \leq n$, and that $F(v_1) = \lambda_1$. Hence we only need to prove that for each $2 \leq i \leq n$

$$F(v_i) \leq \nu_i \lambda_i.$$  

Let $u_1, \ldots, u_n$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \ldots, \lambda_n$, i.e.

$$F(u_i) = \lambda_i, \; \forall \; 1 \leq i \leq n.$$  

Then, by linear independence, for each $2 \leq i \leq n$ at least one of $u_1, \ldots, u_i$ does not belong to the subspace $\text{span}_R \{v_1, \ldots, v_{i-1}\}$, call this vector $u_j$. If the set $v_1, \ldots, v_{i-1}, u_j$ is extendable to a basis of $\Lambda$, then by construction of reduced basis we must have

$$\lambda_i \geq \lambda_j = F(u_j) \geq F(v_i),$$  

and so it implies that $\lambda_i = F(v_i)$, proving (3.6) in this case.

Next assume that the set $v_1, \ldots, v_{i-1}, u_j$ is not extendable to a basis of $\Lambda$. Let $v \in \text{span}_R \{v_1, \ldots, v_{i-1}, u_j\}$ be such that the set $v_1, \ldots, v_{i-1}, v$ is extendable to a basis of $\Lambda$. Then we can write

$$u_j = \sum_{k=1}^{i-1} k_i v_k + \ldots + k_{i-1} v_{i-1} \pm mv,$$

where $k_1, \ldots, k_{i-1}, m \in \mathbb{Z}$, and $m \geq 2$. Indeed, $m \neq 0$ since

$$u_j \notin \text{span}_R \{v_1, \ldots, v_{i-1}\}.$$  

On the other hand, if $m = 1$ then

$$v \in \text{span}_R \{v_1, \ldots, v_{i-1}, u_j\},$$

which would imply that $v_1, \ldots, v_{i-1}, u_j$ is extendable to a basis. Thus $m \geq 2$, and we can write

$$v = \sum_{k=1}^{i-1} \alpha_k v_k + \ldots + \alpha_{i-1} v_{i-1} \pm \frac{1}{m} u_j,$$

where $\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}$. In fact, for each $1 \leq k \leq i - 1$, there exists an integer $l_k$ and a real number $\beta_k$ with $|\beta_k| \leq \frac{1}{2}$ such that

$$\alpha_k = l_k + \beta_k.$$  

Then

$$v = \sum_{k=1}^{i-1} (l_k + \beta_k) v_k \pm \frac{1}{m} u_j = \sum_{k=1}^{i-1} l_k v_k + v',$$
where $v' = \sum_{k=1}^{i-1} \beta_k v_k \pm \frac{1}{m} u_j$. Since $v - v' \in \text{span}_{\mathbb{Z}} \{v_1, \ldots, v_{i-1}\}$, it must be that $v' \in \Lambda$, and the set $v_1, \ldots, v_{i-1}, v'$ is extendable to a basis of $\Lambda$. Then, by definition of $v_i$, we have

$$F(v_i) \leq F(v') \leq \sum_{k=1}^{i-1} F(\beta_k v_k) + F\left(\frac{1}{m} u_j\right) = \sum_{k=1}^{i-1} |\beta_k| F(v_k) + \frac{1}{m} F(u_j) \leq \frac{1}{2} \left(\sum_{k=1}^{i-1} F(v_k) + F(u_j)\right) \leq \frac{1}{2} \left(\sum_{k=1}^{i-1} F(v_k) + \lambda_i\right).$$

Combining this with the previous case, we conclude that

$$F(v_i) \leq \max \left\{ \lambda_i, \frac{1}{2} \left(\sum_{k=1}^{i-1} F(v_k) + \lambda_i\right) \right\}, \ \forall \ 2 \leq i \leq n.$$

Hence we obtain

$$F(v_2) \leq \max \left\{ \lambda_2, \frac{1}{2} (\lambda_1 + \lambda_2) \right\} = \lambda_2,$$

hence $F(v_2) = \lambda_2$. More generally, one can easily deduce (3.6) from (3.7). This finishes the proof.

As a corollary of Theorem 3.2.1, one can easily deduce a bound on the product of diagonal coefficients of reduced positive definite quadratic forms (Problem 3.11).

There are also other reduction procedures for lattice bases, most notably there is a notion of Korkin-Zolotarev reduced basis, which has many applications, for instance in coding theory. In general, depending on particular situation or application one has in mind, one or another reduction may be preferable. The common feature of all reduced bases is that they all contain the shortest nonzero vector of the lattice. One may then ask how to find a Minkowski-reduced basis for a lattice $\Lambda$ with respect to a convex $0$-symmetric set $M$ in $\mathbb{R}^n$? This problem happens to be very difficult in a rather precise sense; in fact, it is a harder version of the Shortest Vector Problem (SVP) that we discussed above.
3.3. Sums of squares

A classical arithmetic problem has to do with representation of integers by positive definite quadratic forms. Indeed, let $Q(X)$ be a positive definite quadratic form in $n$ variables with integer coefficients. Then its values at nonzero integer points are necessarily positive integers. One can then ask which ones? More specifically, given a positive integer $m$, does there exist a point $x \in \mathbb{Z}^n$ such that $Q(x) = m$? If this is the case, we say that $m$ is representable by $Q$.

This question can be interpreted in several different ways. The most obvious one is a question about existence of integer solutions to the equation

$$Q(X) = m. \tag{3.8}$$

Notice that the set of all possible real solutions to (3.8) is actually the surface of an ellipsoid in $\mathbb{R}^n$. Then geometrically our question asks whether this surface contains any integer points? On the other hand, as we know there is a lattice corresponding to $Q(X)$, call it $\Lambda$, so that $Q(X)$ is a norm form corresponding to some choice of basis matrix $A$ of $\Lambda$. Then for any vector $y = Ax \in \Lambda$,

$$\|y\| = Q(x),$$

and so our question is now about the possible integer norm values of vectors in $\Lambda$.

The most natural quadratic form to ask these questions about is the usual Euclidean norm-form, that is the sum of squares: this is the problem we consider in this section. Indeed, results on integers representable as sums of squares go back to at least the work of Pierre de Fermat in 1640, who considered this question in two variables. Namely, Fermat was able to characterize all the integers representable as sums of two integer squares. We start with Fermat’s theorem on representation of primes as sums of two squares: since $2 = 1^2 + 1^2$, we focus on odd primes. There are several known proofs of this result in the literature: we present an elegant proof by an application of Minkowski’s Convex Body Theorem.

**Theorem 3.3.1.** An odd prime number $p$ is representable as a sum of two integer squares if and only if $p \equiv 1 \pmod{4}$.

**Proof.** First notice that for any integer $x$, $x^2$ is congruent to either to 0 or 1 modulo 4. Hence a sum of two integer squares can be congruent to either 0, 1, or 2 modulo 4. Thus a prime that is congruent to 3 modulo 4 cannot be representable as a sum of two squares.

We therefore only need to show that if a prime $p \equiv 1 \pmod{4}$, then there exist $x, y \in \mathbb{Z}$ such that $p = x^2 + y^2$. There exists $m \in \mathbb{Z}$ such that $m^2 \equiv -1 \pmod{p}$ (Problem 3.12). Hence

$$p \mid m^2 + 1.$$ 

Define a lattice

$$\Lambda = \begin{pmatrix} 1 & 0 \\ m & p \end{pmatrix} \mathbb{Z}^2 \subset \mathbb{Z}^2,$$

then $\det(\Lambda) = p$ and any point $u \in \Lambda$ is of the form

$$u = \begin{pmatrix} a \\ am + bp \end{pmatrix}$$

for some integers $a, b$. Then

$$\|u\|^2 = a^2 + (am + bp)^2 = a^2(m^2 + 1) + (2abm + b^2p)p \equiv 0 \pmod{p},$$
hence \( p \mid \|u\|^2 \) for any \( u \in \Lambda \). Let \( \varepsilon > 0 \) and \( B_2(\sqrt{2p} - \varepsilon) \) be a circle of radius \( \sqrt{2p} - \varepsilon \) centered at the origin in the plane. For sufficiently small \( \varepsilon \), the area of \( B_2(\sqrt{2p} - \varepsilon) \) is
\[
\pi(2p - \varepsilon) > 2^2 p = 2^2 \det(\Lambda),
\]
and hence \( B_2(\sqrt{2p} - \varepsilon) \) contains a nonzero point of \( \Lambda \), by Theorem 1.3.2. Let \( u = \left( \begin{array}{c} x \\ y \end{array} \right) \) be this point, then
\[
p \mid \|u\|^2 = x^2 + y^2 \leq 2p - \varepsilon < 2p.
\]
This implies that \( x^2 + y^2 = p \), and so we are done. \( \square \)

We can now deduce a sum of two squares criterion for all the positive integers. For this, we need the following auxiliary lemma.

**Lemma 3.3.2.** Suppose \( a, b \) are integers representable as sums of two squares. Then so is their product \( ab \).

**Proof.** Suppose
\[
a = x^2 + y^2, \quad b = z^2 + t^2
\]
for some \( x, y, z, t \in \mathbb{Z} \). Then
\[
ab = (x^2 + y^2)(z^2 + t^2) = (xz + yt)^2 + (xt - yz)^2.
\]
(3.9)

**Theorem 3.3.3 (Sum of Two Squares).** A positive integer \( m \) is representable as a sum of two integer squares if and only if the prime factors congruent to 3 modulo 4 in its prime factorization occur to an even power.

**Proof.** Let
\[
m = 2^e p_1^{f_1} \cdots p_k^{f_k} q_1^{g_1} \cdots q_n^{g_n}
\]
be the prime decomposition of \( m \), where \( p_1, \ldots, p_k \) are distinct primes \( \equiv 1 \pmod{4} \), \( q_1, \ldots, q_n \) are distinct primes \( \equiv 2 \pmod{4} \), \( e \geq 0 \) and the powers \( f_1, \ldots, f_k, g_1, \ldots, g_n \) are all positive. We then need to show that \( m \) is representable as a sum of two squares if and only if \( g_1, \ldots, g_n \) are all even.

First suppose that all \( g_1, \ldots, g_n \) are all even. Notice that each
\[
q_j^{g_j} = \left( q_j^{g_j/2} \right)^2 + 0^2.
\]
Further, if \( e \) is even, then
\[
2^e = \left( 2^{e/2} \right)^2 + 0^2,
\]
and if \( e \) is odd, then
\[
2^e = 2^{e-1} + 2^{e-1} = \left( 2^{(e-1)/2} \right)^2 + \left( 2^{(e-1)/2} \right)^2.
\]
Finally, each \( p_i \) is representable as a sum of two squares by Theorem 3.3.1. Combining these observations with Lemma 3.3.2, we see that the product \( m \) must be representable as a sum of two squares.
3.3. SUMS OF SQUARES

Now assume $m = x^2 + y^2$ for some $x, y \in \mathbb{Z}$ with $d = \gcd(x, y)$. Let us write $m = a^2b$, where $a, b \in \mathbb{Z}$ with $b$ is squarefree. Then

$$x^2 + y^2 = d^2(x_1^2 + y_1^2) = a^2b,$$

so $d^2 \mid a^2$ and $\gcd(x_1, y_1) = 1$. Suppose some prime $p$ divides $b$, then $p$ must divide $x_1^2 + y_1^2$, i.e.

$$x_1^2 \equiv -y_1^2 \pmod{p}.$$

Suppose $p \equiv 3 \pmod{4}$, then $p - 1 = 4\ell + 2 = 2(2\ell + 1)$, and so

$$x_1^{p-1} = (x_1^2)^{2\ell+1} \equiv (-1)^{2\ell+1}(y_1^2)^{2\ell+1} = -y_1^{p-1} \pmod{p}.$$  

Hence if $p$ divides $x_1$, it must also divide $y_1$, which is not possible since $x_1, y_1$ are relatively prime. Thus $p \nmid x_1, y_1$, in which case Fermat’s Little Theorem implies that $x_1^{p-1}$ and $y_1^{p-1}$ are both congruent to 1 modulo $p$, and hence congruent to each other. This contradicts $(3.11)$, meaning that $p$ cannot be congruent to 3 modulo 4. Hence any odd primes dividing the squarefree part of $m$ must be congruent to 1 modulo 4, so primes congruent to 3 modulo 4 must come to an even power in the factorization $(3.10)$. □

A natural next step in the development of the sum of squares problem is the question of which integers can be represented as sums of three squares? The answer is provided by a theorem of Adrien-Marie Legendre (1797), although (as was observed later) it also follows from an earlier result of Gauss (1796).

**Theorem 3.3.4 (Sum of Three Squares).** A positive integer $m$ is representable as a sum of three integer squares if and only if it is not of the form $m = 4^a(8b + 7)$ for some positive integers $a, b$.

The necessity of the condition $m \neq 4^a(8b + 7)$ is not difficult to see: it follows from the fact that any integer square is either 0, 1, or 4 modulo 8. The sufficiency of this condition is considerably harder; we do not present it here.

Interestingly, the theorem about representing integers as sums of four squares is easier to prove: it was first obtained by Joseph Louis Lagrange in 1770, earlier than Legendre’s theorem. The proof we present here is similar in spirit to our proof of Fermat’s Sum of Two Squares Theorem.

**Theorem 3.3.5 (Sum of Four Squares).** Any positive integer $m$ is representable as a sum of four integer squares.

**Proof.** Similar to the identity $(3.9)$ expressing the product of two sums of two squares as a sum of two squares, there is Euler’s identity for the sum of four squares:

$$\begin{align*}
(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) &= (x_1y_1 - x_2y_2 - x_3y_3 - x_4 y_4)^2 + (x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3)^2 \\
&\quad + (x_1y_3 - x_2y_4 + x_3y_1 + x_4y_2)^2 + (x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1)^2.
\end{align*}$$  

This identity implies that the set of integers representable as sums of four squares is closed under multiplication. Thus we we only need to show that every prime representable this way (obviously, 1 is representable).
Let \( p \) be a prime. There exist some two integers \( a \) and \( b \) so that \( a^2 + b^2 + 1 \) is divisible by \( p \) (Problem 3.13). Define a lattice

\[
\Lambda = \begin{pmatrix} p & 0 & a & b \\ 0 & p & b & -a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \mathbb{Z}^4,
\]

then \( \det(\Lambda) = p^2 \). Let \( u \in \Lambda \), then

\[
\|u\|^2 = (px_1 + ax_3 + bx_4)^2 + (px_2 + bx_3 - ax_4)^2 + x_3^2 + x_4^2
\]

for some \( x_1, \ldots, x_4 \in \mathbb{Z} \). Hence

\[
\|u\|^2 \equiv (x_3^2 + x_4^2)(a^2 + b^2 + 1) \equiv 0 \pmod{p},
\]

thus \( p \mid \|u\|^2 \) for any \( u \in \Lambda \). Let \( \varepsilon > 0 \) and \( B_4(\sqrt{2p - \varepsilon}) \) be a ball of radius \( \sqrt{2p - \varepsilon} \) centered at the origin in \( \mathbb{R}^4 \). For sufficiently small \( \varepsilon \), the volume of \( B_4(\sqrt{2p - \varepsilon}) \) is

\[
\frac{\pi^2}{2} (2p - \varepsilon)^2 > 2^4 p^2 = 2^4 \det(\Lambda),
\]

and hence \( B_4(\sqrt{2p - \varepsilon}) \) contains a nonzero point of \( \Lambda \), by Theorem 1.3.2. Let \( u = (u_1, u_2, u_3, u_4)^T \) be this point, then

\[
p \mid \|u\|^2 = u_1^2 + u_2^2 + u_3^2 + u_4^2 \leq 2p - \varepsilon < 2p.
\]

This implies that \( u_1^2 + u_2^2 + u_3^2 + u_4^2 = p \), and so we are done.

Questions about representation of integers by quadratic forms in general are at the center of an important subarea of number theory, the arithmetic theory of quadratic forms. Ever since the work of Fermat, Lagrange, Legendre and Gauss many mathematicians have studied such representation questions, as well as questions about counting numbers of possible representations. In other words, given an equation of the form (3.8), one can ask:

1. Does it have integer solutions?
2. If so, how many integer solutions does it have?

While we do not address these questions here, we refer the interested reader to the books [HW08] and [MSSW06], where these questions are considered. Some of our arguments in this section followed the exposition of [Cla]. A classical account of the theory of rational quadratic forms can be found in Cassels’ book [Cas78].
3.4. Problems

Problem 3.1. Let $Q(X)$ be a quadratic form in $n$ variables. Prove that
\[
B(X, Y) = \frac{1}{2}(Q(X + Y) - Q(X) - Q(Y))
\]
is a symmetric bilinear form.

Problem 3.2. Let $B$ be an $n \times n$ real symmetric matrix. Let $x_1$ be an eigenvector of $B$ with the corresponding eigenvalue $\lambda_1$ and $\|x_1\| = 1$. Let $U \in O_n(\mathbb{R})$ be a matrix whose columns are an orthonormal basis containing $x_1$ with $x_1$ being the first column. Prove that the matrix $U^\top BU$ is of the form
\[
\begin{pmatrix}
\lambda_1 & 0 & \ldots & 0 \\
0 & a_{11} & \ldots & a_{1(n-1)} \\
\vdots & \vdots & \ddots & \vdots \\
0 & a_{(n-1)1} & \ldots & a_{(n-1)(n-1)}
\end{pmatrix},
\]
where the $(n-1) \times (n-1)$ matrix
\[
A = \begin{pmatrix}
a_{11} & \ldots & a_{1(n-1)} \\
\vdots & \ddots & \vdots \\
a_{(n-1)1} & \ldots & a_{(n-1)(n-1)}
\end{pmatrix}
\]
is also symmetric.

Problem 3.3. Prove that a linear transformation $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is a homomorphism of additive groups, which is an isomorphism if and only if its matrix is nonsingular.

Problem 3.4. Prove that if $\sigma$ is an isomorphism of a symmetric bilinear form $B$, then $\det(\sigma) = \pm 1$. Prove that the set of all isomorphisms of a symmetric bilinear form is a group under matrix multiplication. Hence it must be a subgroup of $GL_n(\mathbb{R})$.

Problem 3.5. Let $B(X, Y)$ be a symmetric bilinear form and $Q(X)$ its associated quadratic form. Prove that the following four conditions are equivalent:

1. $B$ is nonsingular.
2. For every $0 \neq x \in \mathbb{R}^n$, there exists $y \in \mathbb{R}^n$ so that $B(x, y) \neq 0$.
3. For every $0 \neq x \in \mathbb{R}^n$ at least one of the partial derivatives
\[
\frac{\partial Q}{\partial X_i}(x) \neq 0.
\]
4. $Q$ is isometric to a diagonal form with all coefficients nonzero.

Problem 3.6. Prove Sylvester’s Theorem (Theorem 3.1.6), namely that two nonsingular real quadratic forms in $n$ variables are isometric if and only if they have the same signature.
Problem 3.7. Prove that a real quadratic form in \( n \) variables is positive (respectively, negative) definite if and only if it has signature \((n,0)\) (respectively, \((0,n)\)). In particular, a definite form has to be nonsingular.

Problem 3.8. Let \( Q \) be a positive definite real quadratic form in \( n \) variables. Prove that the function \( x \mapsto \sqrt{Q(x)} \) is a norm on \( \mathbb{R}^n \).

Problem 3.9. Prove that every positive definite quadratic form is arithmetically equivalent to a Minkowski reduced form.

Problem 3.10. Let \( B = (b_{ij})_{1 \leq i,j \leq n} \) be the symmetric coefficient matrix of a Minkowski reduced positive definite quadratic form \( Q \). Prove that

\[
0 < b_{11} \leq b_{22} \leq \cdots \leq b_{nn},
\]

and

\[
|2b_{ij}| \leq b_{ii} \quad \forall \ 1 \leq i < j \leq n.
\]

Problem 3.11. Let

\[
Q(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}X_iX_j
\]

be a Minkowski reduced positive definite quadratic form. Prove that

\[
\prod_{i=1}^{n} b_{ii} \leq \frac{4^n}{\omega_n} \left( \frac{3}{2} \right)^{\frac{(n-1)(n-2)}{2}} \det(Q),
\]

where \( \omega_n \) is the volume of a unit ball in \( \mathbb{R}^n \), which is given by (2.1).

(Hint: Let \( \Lambda = \mathbb{Z}^n \), and let \( M = \{ x \in \mathbb{R}^n : \sqrt{Q(x)} \leq 1 \} \); then apply Theorem 3.2.1.)

Problem 3.12. Let \( p \) be a prime congruent to 1 modulo 4. Use Euler’s Criterion to prove that there exists \( m \in \mathbb{Z} \) such that \( m^2 \equiv -1 \pmod{p} \).

Problem 3.13. Let \( p \) be a prime. Prove that there exist some two integers \( a \) and \( b \) so that \( a^2 + b^2 + 1 \) is divisible by \( p \).
CHAPTER 4

Diophantine Approximation

4.1. Real and rational numbers

Diophantine approximation aims to quantify the quality of approximation of real numbers by rationals. The set $\mathbb{Q}$ of rational numbers can be defined as the set of equivalence classes of integer pairs $(a, b) \in \mathbb{Z}^2$ under the relation

\[(a, b) \sim (c, d) \iff ad = bc.\]

We can then construct real numbers as equivalence classes of rational Cauchy sequences under the relation that two sequences are equivalent whenever they converge to the same limit. In terms of decimal expansion we can define a real number to be a power series

$$\sum_{k=0}^{\infty} a_k 10^{-k},$$

where the coefficients $a_0, a_1, \ldots$ are integers in the interval $[-9, 9]$, either all non-negative or all nonpositive. As indicated in Problem 4.1, each such power series converges.

Now, rational numbers are those power series for which either all but finitely many $a_k$ are zero, or those for which the sequence of coefficients $\{a_0, a_1, \ldots\}$ is periodic. From this description it is clear that most real numbers must be irrational, but can this statement be made more precise? In this chapter we will explore the relationship between rational and all real numbers in some detail, drawing precise conclusions.

The first observation we make is that although rationals are sparse among the reals, it is always possible to find a rational number as close as we want to a given real number.

**Theorem 4.1.1.** The set of rational numbers $\mathbb{Q}$ is dense inside of the set of real number $\mathbb{R}$, i.e. if $x < y \in \mathbb{R}$, then there exists $z \in \mathbb{Q}$ such that

$$x < z < y.$$
Let \( z = \frac{m}{n} \in \mathbb{Q} \), and this finishes the proof.

Theorem 4.1.1 implies that any real number can be approximated arbitrarily well by rational numbers. In the next section we show that, while this is true, the number of rationals is still incomparably smaller than the number of all reals in a certain well defined sense.
4.2. Algebraic and transcendental numbers

We start out with definitions and basic properties of algebraic and transcendental numbers.

**Definition 4.2.1.** A complex number \( \alpha \) is called **algebraic** if there exists a nonzero polynomial \( p(x) \) with integer coefficients such that \( p(\alpha) = 0 \). If \( \alpha \) is not algebraic, it is called **transcendental**.

In other words, transcendental numbers are complex numbers that do not satisfy any polynomial equation with integer coefficients. We will write \( \mathbb{A} \) for the set of all algebraic numbers and \( \mathbb{T} := \mathbb{C} \setminus \mathbb{A} \) for the set of all transcendental numbers.

Examples of algebraic numbers are easy to construct. In fact, it is easily seen that every rational number \( \frac{m}{n} \) is algebraic: it is the root of polynomial \( p(x) = nx - m \). More generally, any number of the form \( \left( \frac{m}{n} \right)^{1/k} \), where \( m, n \) are integers, \( n \neq 0 \), and \( k \) a positive integer is also algebraic: it is a root of the polynomial \( p(x) = nx^k - m \). Notice that this example includes such instances as \( \sqrt{2} \), \( i = \sqrt{-1} \), and many others. These examples and the ease with which they can be constructed may give an impression that most complex numbers are algebraic. In fact, this is not true. Our first goal is to make this idea rigorous.

First let us introduce some additional notation. Recall that we write \( \mathbb{Z}[x] \) for the ring of all polynomials with integer coefficients. We think of constants as polynomials of degree 0, and hence \( \mathbb{Z} \subset \mathbb{Z}[x] \). The **degree** of an algebraic number \( \alpha \) is defined as

\[
\deg(\alpha) := \min \{ \deg(f(x)) : f(x) \in \mathbb{Z}[x], f(\alpha) = 0 \}.
\]

Let \( d = \deg(\alpha) \) and let \( f(x) = \sum_{m=0}^{d} a_m x^m \in \mathbb{Z}[x] \) be a polynomial of degree \( d \) such that \( f(\alpha) = 0 \), \( \gcd(a_0, \ldots, a_d) = 1 \), and \( a_d > 0 \). By Problem 4.2, this polynomial is unique for each \( \alpha \in \mathbb{A} \): it is called the **minimal polynomial** of \( \alpha \), denoted by \( m_\alpha(x) \).

A polynomial \( p(x) \in \mathbb{Z}[x] \) is called **irreducible** if whenever \( p(x) = f(x)g(x) \) for some \( f(x), g(x) \in \mathbb{Z}[x] \) then either \( f(x) \) or \( g(x) \) is equal to \( \pm 1 \).

**Definition 4.2.2.** A set \( S \) is called **countable** if there exists a bijective (i.e., one-to-one and onto) map \( f : \mathbb{N} \to S \).

**Lemma 4.2.1.** Let \( S_1, S_2, \ldots \) be a collection of finite sets. Then their union

\[
S = \bigcup_{n=1}^{\infty} S_n
\]

is countable.

**Proof.** For each \( n \geq 1 \), let \( a_n \) be the cardinality of \( S_n \), and write

\[
S_n = \{ x_{n1}, \ldots, x_{na_n} \}.
\]

Then we can write

\[
S = \{ x_{11}, \ldots, x_{1a_1}, x_{21}, \ldots, x_{2a_2}, \ldots \}.
\]

Let \( y_m \) be the \( m \)-th element of \( S \) with respect to the above ordering, i.e. \( y_m = x_{nj} \) for some \( n \) and \( j \) such that

\[
a_1 + \cdots + a_{n-1} + j = m.
\]

Then define \( f : \mathbb{N} \to S \) by \( f(m) = y_m \). This map is clearly a bijection, and hence \( S \) is countable. \( \square \)
Lemma 4.2.2. The set \( \mathbb{N} \times \mathbb{N} \) is countable.

Proof. Notice that
\[
\mathbb{N} \times \mathbb{N} = \{(m, n) : m, n \in \mathbb{N}\} = \{(m, n) : m, n \in \mathbb{N}, m \leq n\} \cup \{(m, n) : m, n \in \mathbb{N}, m > n\} = \left( \bigcup_{n=1}^{\infty} \{(m, n) : m \leq n\} \right) \cup \left( \bigcup_{m=1}^{\infty} \{(m, n) : n < m\} \right),
\]
which is a (countable) union of finite sets, and hence it is countable by Lemma 4.2.1 above.

Lemma 4.2.3. A countable union of countable sets is countable.

Proof. Let \( S_1, S_2, \ldots \) be countable sets, say
\[
S_n = \{x_{n1}, x_{n2}, \ldots\},
\]
and let
\[
S = \bigcup_{n=1}^{\infty} S_n.
\]
Then notice that there is a bijection \( f : \mathbb{N} \times \mathbb{N} \to S \), given by \( f(n, m) = x_{nm} \). By Lemma 4.2.2, \( \mathbb{N} \times \mathbb{N} \) is countable, i.e. there exists a bijection \( g : \mathbb{N} \to \mathbb{N} \times \mathbb{N} \). Since a composition of two bijections \( f \circ g : \mathbb{N} \to S \) is again a bijection, we conclude that \( S \) is countable.

Lemma 4.2.4. Let \( m \geq 1 \). The set
\[
\mathbb{Z}^m := \{a = (a_1, \ldots, a_m) : a_1, \ldots, a_m \in \mathbb{Z}\}
\]
is countable.

Proof. We argue by induction on \( m \). First suppose that \( m = 1 \), then the set
\[
\mathbb{Z} = \mathbb{N} \cup -\mathbb{N} \cup \{0\},
\]
where \( -\mathbb{N} = \{-x : x \in \mathbb{N}\} \). This is a union of two countable sets and one finite set, hence it is countable. Now suppose that the statement of the lemma is true for \( m = d - 1 \). We prove it for \( m = d \). Notice that
\[
\mathbb{Z}^d = \left( \bigcup_{a \in \mathbb{N}_0} \{(x, a) : x \in \mathbb{Z}^{d-1}\} \right) \cup \left( \bigcup_{a \in \mathbb{N}} \{(x, -a) : x \in \mathbb{Z}^{d-1}\} \right),
\]
Each set like \( \{(x, a) : x \in \mathbb{Z}^{d-1}\} \) or \( \{(x, -a) : x \in \mathbb{Z}^{d-1}\} \) for \( a \in \mathbb{N} \) is in bijective correspondence with \( \mathbb{Z}^{d-1} \), and hence is countable by induction hypothesis. Therefore \( \mathbb{Z}^d \) is a countable union of countable sets, and hence is countable by Lemma 4.2.3.

Remark 4.2.1. One can use Lemma 4.2.4 to deduce that \( \mathbb{Q} \) is a countable set. Indeed, rational numbers are constructed as the set of equivalence classes of the subset \( \mathbb{Z}^2 \) under the specified equivalence relation (4.1). Identifying these equivalence classes with some choice of their representatives, we can view \( \mathbb{Q} \) as a subset of \( \mathbb{Z}^2 \). Lemma 4.2.4 implies that \( \mathbb{Z}^2 \) is countable, and then Problem 4.4 guarantees that \( \mathbb{Q} \) is countable.

We will now prove a much stronger fact, namely countability of the set of all algebraic numbers, from which countability of \( \mathbb{Q} \) follows yet again by Problem 4.4.
Theorem 4.2.5. The set $\mathbb{A}$ of algebraic numbers is countable.

Proof. Notice that each $\alpha \in \mathbb{A}$ is a root of some polynomial in $\mathbb{Z}[x]$. Furthermore, each polynomial $p(x) \in \mathbb{Z}[x]$ has finitely many roots. In fact, the Fundamental Theorem of Algebra (presented in Appendix B) guarantees that any polynomial with coefficients in $\mathbb{C}$ has $d$ roots in $\mathbb{C}$, counted with multiplicity, where $d$ is its degree. For each $p(x) \in \mathbb{Z}[x]$, let $R_p$ be the set of all roots of $p(x)$. Then

$$\mathbb{A} = \bigcup_{p(x) \in \mathbb{Z}[x]} R_p.$$  

This union is not disjoint, i.e. roots may be repeated. Hence, if we just think of this union as a list of elements with repetition, then $\mathbb{A}$ is formally a subset of $\bigcup_{p(x) \in \mathbb{Z}[x]} R_p$. Now notice that each polynomial

$$p(x) = \sum_{n=0}^{d} a_n x^n \in \mathbb{Z}[x]$$

can be identified with its vector of coefficients $(a_0, \ldots, a_d) \in \mathbb{Z}^{d+1}$, where $d = \deg(p(x))$. This defines a bijection between $\mathbb{Z}[x]$ and the set $\bigcup_{n \in \mathbb{N}_0} \mathbb{Z}^{d+1}$, which is a countable union of countable sets, hence is countable. Therefore $\bigcup_{p(x) \in \mathbb{Z}[x]} R_p$ is a countable union of finite sets, hence is countable, and so its subset $\mathbb{A}$ is also countable.

Remark 4.2.2. In fact, we could rephrase the proof Theorem 4.2.5 in terms of just irreducible polynomials. In other words, there is a bijection between $\mathbb{A}$ and the disjoint union of sets of roots of all irreducible polynomials in $\mathbb{Z}[x]$. Since the set of irreducible polynomials is an infinite subset of the countable set $\mathbb{Z}[x]$, it is itself countable, hence we are done.

In contrast, let us consider the set of all real numbers.

Theorem 4.2.6. The set $\mathbb{R}$ of all real numbers is uncountable.

Proof. Assume that $\mathbb{R}$ is countable. Then there exists some bijection $f : \mathbb{N} \to \mathbb{R}$. Let us write $x_n := f(n)$ for each $n \in \mathbb{N}$, so the image of $f$ is the sequence $(x_n)_{n \in \mathbb{N}}$ of distinct real numbers, which is supposed to be equal to all of $\mathbb{R}$. We will reach a contradiction by showing that every sequence $(x_n)_{n \in \mathbb{N}}$ of distinct real numbers misses at least one $x \in \mathbb{R}$.

Indeed, let $(x_n)_{n \in \mathbb{N}}$ be such a sequence. We define a nested family of intervals as follows. Let $a_1 := \min\{x_1, x_2\}$ and $b_1 := \max\{x_1, x_2\}$. Since the elements of our sequence are all distinct, $a_1 < b_1$, and hence $I_1 := [a_1, b_1]$ is an interval, not a singleton. If $I_1$ contains only finitely many $x_n$’s, then pick some $x \in I_1$ which is not one of these numbers (by Problem 4.5, such $x$ must exist), and we are done. Then assume $I_1$ contains infinitely many $x_n$’s. Let $y$ and $z$ be the first two such elements, with respect to index, in the interior of $I_1$ and let $a_2 := \min\{y, z\}$, $b_2 := \max\{y, z\}$ so $a_2 < b_2$ and $I_2 := [a_2, b_2]$ is again an interval with non-empty interior such that $I_2 \subseteq I_1$. Continue in the same manner to obtain a nested sequence of intervals:

$$\cdots \subseteq I_1 \subseteq I_{n-1} \subseteq \cdots \subseteq I_2 \subseteq I_1,$$

where each $I_n := [a_n, b_n]$ with $a_n < b_n$. Then notice that

$$a_1 < a_2 < \cdots < a_{n-1} < a_n < \cdots < b_n < b_{n-1} < \cdots < b_2 < b_1.$$
Therefore \((a_n)_{n \in \mathbb{N}}\) (respectively, \((b_n)_{n \in \mathbb{N}}\)) is a monotone increasing (respectively, decreasing) sequence, which is bounded from above (respectively, below). By the Monotone Convergence Theorem (recall from Calculus), these sequences have limits, let us write

\[ A := \lim_{n \to \infty} a_n, \quad B := \lim_{n \to \infty} b_n. \]

It is clear that \(A \leq B\), so the closed interval \(I = [A, B]\) is not empty. Let \(h \in I\), then \(h \neq a_n, b_n\) for any \(n \in \mathbb{N}\). In fact, we will show that \(h \neq x_n\) for any \(n \in \mathbb{N}\).

Suppose that \(h = x_k\) for some \(k \in \mathbb{N}\), so there are finitely many points in the sequence \((x_n)_{n \in \mathbb{N}}\) before \(h\) occurs, and hence only finitely many \(a_n\)'s preceding \(h\). Let \(a_d\) be the last element in the sequence \((a_n)_{n \in \mathbb{N}}\) preceding \(h\). Since \(h\) cannot be equal to \(a_d\), \(a_d < h\), i.e. \(h\) is in the interior of \(I_d\). Since it is contained in the limiting interval \(I\), it must be contained in \(I_{d+1} = [a_{d+1}, b_{d+1}]\) by our construction of the intervals. But this means that \(a_d < a_{d+1} < h\), which contradicts our choice of \(a_d\).

This shows that \(h\) is not an element of the sequence \((x_n)_{n \in \mathbb{N}}\), and hence at least one real number is not in this sequence. This means that \(\mathbb{R}\) cannot be countable. \(\square\)

**Remark 4.2.3.** The fact of uncountability of reals was first established by Georg Cantor in 1874. In fact, Cantor presented at least three different proofs of this fact, including his famous diagonal argument (1891). Our proof of Theorem 4.2.6 above follows Cantor’s first argument (1874).

Since \(\mathbb{R} \subset \mathbb{C}\), we conclude that \(\mathbb{C}\) is also uncountable, by Problem 4.4. Now recall that \(\mathbb{C} = \mathbb{A} \cup \mathbb{T}\), and \(\mathbb{A}\) is countable. This means that \(\mathbb{T}\), the set of transcendental numbers, is uncountable. Loosely speaking this means, that most complex numbers are in fact transcendental. Ironically, while constructing algebraic numbers is quite straightforward, as seen above, it is not at all easy to construct a transcendental number. Indeed, suppose we take a complex number \(\alpha\). To prove that it is algebraic, we can find its minimal polynomial \(m_\alpha(x) \in \mathbb{Z}[x]\). Although this may be somewhat laborious, there are standard techniques in algebraic number theory that allow for such a construction. On the other hand, to prove that \(\alpha\) is transcendental we would need to establish that \(\alpha\) is not a root of any polynomial in \(\mathbb{Z}[x]\). This kind of fact clearly requires some sort of indirect argument, which is the reason why it took mathematicians until mid-19th century to construct the first transcendental number. This construction, by Joseph Liouville, used the recently developed tools in the area of Diophantine approximation. It is our next goal to develop the necessary tools and to present Liouville’s construction. Our exposition in the next three sections follows the classical text of W. M. Schmidt [Sch91].
4.3. Dirichlet’s Theorem

Since rationals are dense within reals, we can always approximate a real number with rationals. For many purposes, however, we may want to control how “complicated” the rational numbers we use for such approximations are, i.e. we may want to bound the size of their denominators. This is the starting point of the theory of Diophantine approximation. The first result in this direction dates back to Dirichlet, and is proved with the use of Dirichlet’s box principle (also known in combinatorics as the pigeonhole principle); in fact, this is the theorem to which this principle owes its name.

**Theorem 4.3.1 (Dirichlet, (1842)).** Let \( \alpha \in \mathbb{R} \), and let \( Q \in \mathbb{Z}_{>0} \). There exist relatively prime integers \( p,q \) with \( 1 \leq q \leq Q \) such that

\[
\left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q(Q+1)}.
\]

Moreover, if \( \alpha \) is irrational, then there are infinitely many rational numbers \( \frac{p}{q} \) such that

\[
\left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q^2}.
\]

**Proof.** If \( \alpha \) is a rational number with denominator \( \leq Q \), there is nothing to prove. Hence we will assume that either \( \alpha \) is irrational, or it is rational with denominator \( >Q \). Notice that

\[
[0,1) = \bigcup_{i=1}^{Q+1} \left( \frac{i-1}{Q+1}, \frac{i}{Q+1} \right).
\]

Consider the numbers \( \{l\alpha\}, 1 \leq l \leq Q+1 \), where \( \{ \} \) denotes the fractional part function, i.e. \( \{x\} = x - \lfloor x \rfloor \). These numbers lie in the interval \([0,1)\) and are distinct. Indeed, suppose that \( \{l\alpha\} = \{m\alpha\} \) for some \( 1 \leq l < m \leq Q+1 \), then \( m\alpha - l\alpha \) is an integer, say

\[
m\alpha - l\alpha = (m-l)\alpha = k \in \mathbb{Z},
\]

and so \( \alpha = k/(m-l) \), where \( m-l \leq (Q+1) - 1 = Q \), which contradicts our assumption.

**Case 1.** Suppose that each subinterval \( \left[ \frac{i-1}{Q+1}, \frac{i}{Q+1} \right) \) contains one of the numbers \( \{l\alpha\}, 1 \leq l \leq Q+1 \). In particular, subintervals \( \left[ 0, \frac{1}{Q+1} \right) \) and \( \left[ \frac{Q}{Q+1}, 1 \right) \) contain such points, so at least one of them must contain some \( \{l\alpha\} \) with \( 1 \leq l \leq Q \). Therefore, either

\[
|l\alpha - \lfloor l\alpha \rfloor| \leq \frac{1}{Q+1},
\]

or

\[
|l\alpha - \lfloor l\alpha \rfloor - 1| \leq \frac{1}{Q+1}.
\]

This means that there exists an integer \( 1 \leq l \leq Q \) and an integer \( m \) equal to either \( \lfloor l\alpha \rfloor \) or \( \lfloor l\alpha \rfloor - 1 \), depending on whether (4.4) or (4.5) holds, such that

\[
|l\alpha - m| \leq \frac{1}{Q+1}.
\]
Let \( d = \gcd(l, m) \), and let \( p = \frac{m}{d} \) and \( q = \frac{l}{d} \), then
\[
|qd\alpha - pd| \leq \frac{1}{Q + 1},
\]
meaning that
\[
\left| \alpha - \frac{p}{q} \right| \leq \frac{1}{qd(Q + 1)} \leq \frac{1}{q(Q + 1)},
\]
proving (4.2) in this case.

**Case 2.** Now assume that one of the subintervals \( \left[ \frac{i-1}{Q+1}, \frac{i}{Q+1} \right) \) for some \( 1 \leq i \leq Q + 1 \) does not contain any of the numbers \( \{l\alpha\} \), \( 1 \leq l \leq Q + 1 \). Since there are \( Q + 1 \) such numbers and \( Q + 1 \) subintervals, one of the subintervals must contain two such numbers, say \( \left[ \frac{j-1}{Q+1}, \frac{j}{Q+1} \right) \) for some \( 1 \leq j \leq Q+1 \) contains \( \{l\alpha\} \) and \( \{m\alpha\} \) for some \( 1 \leq l < m \leq Q + 1 \). Therefore
\[
|(m\alpha - \lfloor m\alpha \rfloor) - (l\alpha - \lfloor l\alpha \rfloor)| = |(m - l)\alpha - (\lfloor m\alpha \rfloor - \lfloor l\alpha \rfloor)| \leq \frac{1}{Q + 1}.
\]
Once again, let \( d = \gcd((m - l), (\lfloor m\alpha \rfloor - \lfloor l\alpha \rfloor)) \), and let \( p = \frac{\lfloor m\alpha \rfloor - \lfloor l\alpha \rfloor}{d} \) and \( q = \frac{m - l}{d} \), and so in the same way as above we obtain (4.2).

We can now derive (4.3) from (4.2): since \( q \leq Q \),
\[
(4.6) \quad \left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q(Q + 1)} < \frac{1}{q^2}.
\]
Now suppose that there are only finitely many rationals that satisfy (4.3), call them
\[
\frac{p_1}{q_1}, \ldots, \frac{p_k}{q_k}.
\]
Let
\[
\delta = \min_{1 \leq i \leq k} \left| \alpha - \frac{p_i}{q_i} \right|,
\]
then \( \delta > 0 \), since \( \alpha \) is irrational. Let \( Q \in \mathbb{Z}_{>0} \) be such that
\[
\frac{1}{Q} < \delta.
\]
By (4.6), there must exist \( \frac{p}{q} \) with \( 1 \leq q \leq Q \) such that
\[
\left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q(Q + 1)} < \delta,
\]
hence \( \frac{p}{q} \notin \left\{ \frac{p_1}{q_1}, \ldots, \frac{p_k}{q_k} \right\} \), which is a contradiction. Thus there must be infinitely many such rationals.

**Remark 4.3.1.** Notice that the argument that derives (4.3) from (4.2) is very similar to Euclid’s proof of the infinitude of primes.

We also present an alternate proof of Dirichlet’s inequality (4.3), which is inspired by the geometry of numbers approach and uses Minkowski’s Linear Forms Theorem. Our exposition of this proof follows [Cla].
Minkowski-style proof of Dirichlet’s theorem. With notation as in the statement of Theorem 4.3.1, let us define two binary linear forms:

\[ L_1(x, y) = x - \alpha y, \quad L_2(x, y) = y. \]

Coefficients of these forms are given by the rows of the \( 2 \times 2 \) matrix

\[ B = \begin{pmatrix} 1 & -\alpha \\ 0 & 1 \end{pmatrix} \]

with \( \det(B) = 1 \). Let \( c_1, c_2 \in \mathbb{R} \) be such that \( c_1c_2 = 1 \), then by Theorem 1.3.3 there exists a nonzero point \((p, q) \in \mathbb{Z}^2\) such that \(|L_1(p, q)| \leq c_1, |L_2(p, q)| \leq c_2\), i.e.

\[ |p - \alpha q| \leq \frac{1}{c_2}, \quad |q| \leq c_2. \]

If we take \( c_2 > 1 \), the first of these inequalities implies that \( q \neq 0 \): otherwise \( p \) would have to be 0 too, contradicting the fact that \((p, q)\) is a nonzero lattice point.

Let \( h \in (0, 1) \) and set \( c_2 = Q + h \), then for any such \( h \) there exist \( p, q \in \mathbb{Z} \) with \( q \leq Q + h \) (hence \( \leq Q \) since \( q \) is an integer) such that

\[ |p - \alpha q| \leq \frac{1}{Q + h}. \]

Since this inequality holds for any \( h \), and there are only finitely many points \((p, q) \in \mathbb{Z}^2\) with \( |q| \leq Q \) satisfying this, there must in fact exist \((p, q) \in \mathbb{Z}^2\) such that

\[ |p - \alpha q| \leq \frac{1}{Q + 1} \quad \text{and} \quad |q| \leq Q. \]

Dividing through by \( q \) completes the proof. \( \Box \)

Hurwitz (1891) improved Dirichlet’s bound (4.3) slightly by showing that for any irrational \( \alpha \in \mathbb{R} \) there exist infinitely many distinct rational numbers \( \frac{p}{q} \) such that

\[ \left| \alpha - \frac{p}{q} \right| < \frac{1}{\sqrt{5}q^2}. \]

We will now show that in a certain sense (4.7) is best possible.

Lemma 4.3.2. Let \( \alpha \in \mathbb{R} \) be a quadratic irrational satisfying \( f(\alpha) = 0 \), where

\[ f(x) = ax^2 + bx + c \]

with \( a, b, c \in \mathbb{Z} \) and \( a > 0 \). Write \( D = b^2 - 4ac \) for the discriminant of \( f \). Then for any real number \( A > \sqrt{D} \), there are only finitely many rationals \( \frac{p}{q} \) such that

\[ \left| \alpha - \frac{p}{q} \right| < \frac{1}{Aq^2}. \]

Proof. We know that \( \alpha \) is one of the roots of \( f(x) \), then let \( \beta \) be the other one, i.e.

\[ f(x) = a(x - \alpha)(x - \beta) = ax^2 - a(\alpha + \beta)x + a\alpha\beta, \]

meaning that \( b = a(\alpha + \beta) \) and \( c = a\alpha\beta \). Therefore

\[ D = b^2 - 4ac = a^2(\alpha - \beta)^2. \]

Now suppose that for some \( \frac{p}{q} \in \mathbb{Q} \), (4.8) holds. Notice that since \( f(x) \) is a quadratic polynomial with irrational roots, then

\[ 0 \neq \left| f \left( \frac{p}{q} \right) \right| = \left| \frac{ap^2 + bpq + cq^2}{q^2} \right| \geq \frac{1}{q^2}. \]
since $0 \neq ap^2 + bpq + cq^2 \in \mathbb{Z}$, hence $|ap^2 + bpq + cq^2| \geq 1$. Therefore
\[
\frac{1}{q^2} \leq |f\left(\frac{p}{q}\right)| = a \left| \alpha - \frac{p}{q} \right| \left| \beta - \frac{p}{q} \right| \\
< \frac{a}{Aq^2} \left| \beta - \frac{p}{q} \right| = \frac{a}{Aq^2} \left( \left| \alpha - \frac{p}{q} \right| + (\beta - \alpha) \right) \\
\leq \frac{a}{Aq^2} \left| \alpha - \frac{p}{q} \right| + \frac{a}{Aq^2} |\beta - \alpha| < \frac{a}{A^2q^2} + \frac{\sqrt{D}}{Aq^2},
\]
and subtracting $\frac{\sqrt{D}}{Aq^2}$ from both sides of the above inequality implies
\[
\frac{1}{q^2} \left(1 - \frac{\sqrt{D}}{A}\right) < \frac{a}{A^2q^2}.
\]
The left hand side of this inequality is not 0 since $A > \sqrt{D}$, and hence
\[
q^2 < \frac{a}{A(A - \sqrt{D})}.
\]
This implies that there are only finitely many possibilities for the denominator $q$, but for each such $q$ there can be only finitely many $p$ so that (4.8) holds. This completes the proof. \[\square\]

**Remark 4.3.2.** Let $\alpha = \frac{1 + \sqrt{5}}{2}$, then the corresponding polynomial
\[
f(x) = x^2 - x - 1,
\]
and its discriminant is $D = 5$. By Lemma 4.3.2, if $A > \sqrt{5}$ then there are only finitely many $\frac{p}{q} \in \mathbb{Q}$ such that
\[
\left| \alpha - \frac{p}{q} \right| < \frac{1}{Aq^2},
\]
which proves that Hurwitz’s bound (4.7) is best possible.
4.4. Liouville’s theorem and construction of a transcendental number

More generally, for every quadratic irrational \( \alpha \) there exists a constant \( C(\alpha) > 0 \) such that for any \( \frac{p}{q} \in \mathbb{Q} \)

\[
\left| \alpha - \frac{p}{q} \right| \geq \frac{C(\alpha)}{q^2}.
\]

In other words, quadratic irrationals are badly approximable.

**Definition 4.4.1.** An irrational number \( \alpha \) is called *badly approximable* if there exists a positive real constant \( C(\alpha) \) such that (4.9) holds for any \( \frac{p}{q} \in \mathbb{Q} \).

As can be expected after the above discussion, algebraic numbers although are not necessarily badly approximable, are certainly “worth” approximable than transcendental can be. This principle was first observed by Liouville in 1844.

**Theorem 4.4.1 (Liouville).** Let \( \alpha \in \mathbb{R} \) be an algebraic number of degree \( d = \deg(f) \geq 2 \), where \( f(x) \in \mathbb{Z}[x] \) is the minimal polynomial of \( \alpha \) over \( \mathbb{Q} \). Then there exists a positive real constant \( C(\alpha) \) such that for any \( \frac{p}{q} \in \mathbb{Q} \)

\[
\left| \alpha - \frac{p}{q} \right| \geq \frac{C(\alpha)}{q^d}.
\]

**Proof.** Let 

\[ f(x) = \sum_{i=0}^{d} a_i x^i \in \mathbb{Z}[x]. \]

Then, since \( d \geq 2 \) means that \( \alpha \) is irrational, for each \( \frac{p}{q} \in \mathbb{Q} \) we have

\[ 0 \neq q^d f \left( \frac{p}{q} \right) = \sum_{i=0}^{d} a_i p^i q^{d-i} \in \mathbb{Z}. \]

We can assume of course that \( \left| \alpha - \frac{p}{q} \right| \leq 1 \). Then, since \( f(\alpha) = 0 \),

\[
1 \leq q^d \left| f \left( \frac{p}{q} \right) \right| = q^d \left| f(\alpha) - f \left( \frac{p}{q} \right) \right| = q^d \left| \int_{p/q}^{\alpha} f'(u) \, du \right| \\
\leq q^d \left| \alpha - \frac{p}{q} \right| \max \{ f'(u) : |\alpha - u| \leq 1 \}.
\]

Then pick \( C(\alpha) = (\max \{ f'(u) : |\alpha - u| \leq 1 \})^{-1} \), and the theorem follows. \( \square \)

Liouville used his theorem to construct the first known example of a transcendental number.

**Corollary 4.4.2 (Liouville).** The number

\[ \alpha = \sum_{n=1}^{\infty} \frac{1}{a^n} \]

is transcendental for any integer \( a \geq 2 \).

**Proof.** Let \( a > 1 \). For every \( k \in \mathbb{Z}_{>0} \), let

\[ p_k = a^{k!} \sum_{n=1}^{k} \frac{1}{a^n}, \quad q_k = a^{k!} \in \mathbb{Z}. \]
Then
\[ |\alpha - \frac{p_k}{q_k}| = \sum_{n=k+1}^{\infty} \frac{1}{a^n!} = \frac{1}{a(k+1)!} \sum_{n=k+1}^{\infty} \frac{a^{(k+1)!}}{n!} < \frac{1}{a(k+1)!} \sum_{n=0}^{\infty} \frac{1}{a^n}. \]

Clearly \( \sum_{n=0}^{\infty} \frac{1}{a^n} \) is a convergent series, so let
\[ C = \sum_{n=0}^{\infty} \frac{1}{a^n}, \]
and then we have
\[ (4.11) \quad |\alpha - \frac{p_k}{q_k}| < \frac{C}{a(k+1)!} = \frac{C}{q_k^{(k+1)!}} < \frac{C}{q_k}. \]

Suppose that \( \alpha \) is rational, say \( \alpha = c/d \) for some \( c, d, \in \mathbb{Z} \). Then (4.11) implies that
\[ |cq_k - dp_k| < \frac{C}{k^{k-1}} \]
for infinitely many \( p_k/q_k \) as above. The expression \( \frac{C}{q_k} \) is < 1 for all large enough \( q_k \). On the other hand, \( |cq_k - dp_k| \) is a nonnegative integer, which can be 0 for at most one \( p_k/q_k \); hence \( |cq_k - dp_k| \geq 1 \) for infinitely many \( p_k/q_k \). This is a contradiction, and so \( \alpha \) cannot be rational.

Now suppose that \( \alpha \) is algebraic of degree \( d \). Then, by Theorem 4.4.1, there exists a constant \( C(\alpha) \) such that
\[ |\alpha - \frac{p_k}{q_k}| \geq \frac{C(\alpha)}{q_k^d}, \]
for every \( k \in \mathbb{Z}_{>0} \). However, if we take \( k \) large enough so that
\[ \frac{C}{q_k^d} < \frac{C(\alpha)}{q_k^d}, \]
then (4.11) implies a contradiction; more specifically, we just need to take \( k \) large enough so that
\[ k!(k - d) > \frac{\ln C - \ln C(\alpha)}{\ln a}. \]

This completes the proof. \( \square \)

Remark 4.4.1. Numbers that can be proved to be transcendental using Liouville’s theorem are called Liouville numbers; they form a rather small set. In particular, \( e \) and \( \pi \) (which are transcendental) are not Liouville numbers, and neither are most transcendental numbers.
4.5. Roth’s theorem

Theorem 4.4.1 implies that if $\alpha \in \mathbb{R}$ is an algebraic number of degree $d \geq 2$ and $\mu > d$, then there are only finitely many $\frac{p}{q} \in \mathbb{Q}$ with $\gcd(p, q) = 1$ such that

$$
\left| \alpha - \frac{p}{q} \right| < \frac{1}{q^\mu}.
$$

(4.12)

Indeed, suppose there were infinitely many rational numbers for which (4.12) holds. Let $C(\alpha)$ be the constant guaranteed by Theorem 4.4.1. Let $Q$ be an integer so that $\frac{1}{Q^{\mu-2}} > C(\alpha)$. Clearly there can be only finitely many $\frac{p}{q}$ with $\gcd(p, q) = 1$ for which (4.12) holds with $q \leq Q$, hence there must be infinitely many such rationals with $q > Q$. Suppose $\frac{p}{q}$ is one of them, then

$$
\left| \alpha - \frac{p}{q} \right| < \frac{1}{q^\mu} < \frac{1}{Q^{\mu-2}q^d} < \frac{C(\alpha)}{q^d},
$$

which contradicts (4.10). This proves finiteness of the number of solutions for (4.12).

For an algebraic number $\alpha$ of degree $d \geq 2$, what is the smallest possible $\mu$ for which (4.12) will have only finitely many solutions? Combining the discussion above with Dirichlet’s theorem (Theorem 4.3.1), we see that

$$
2 \leq \mu \leq d + \delta,
$$

for any $\delta > 0$. In 1908 Thue proved that $\mu \leq \frac{d+2}{2} + \delta$; in 1921 Siegel proved that $\mu \leq 2\sqrt{d} + \delta$. Dyson (1947) and Gelfond (1952) proved that $\mu \leq \sqrt{2d} + \delta$. The major breakthrough came with the famous theorem of Roth (1955) [Rot55], for which he received a Fields medal in 1958.

**Theorem 4.5.1 (Roth).** Let $\alpha \in \mathbb{R}$ be an algebraic number. For any $\delta > 0$, there are only finitely many rationals $\frac{p}{q}$ with $\gcd(p, q) = 1$ such that

$$
\left| \alpha - \frac{p}{q} \right| < \frac{1}{q^{2+\delta}}.
$$

**Remark 4.5.1.** Dirichlet’s theorem shows that Roth’s theorem is best possible, i.e. the exponent on $q$ in the upper bound cannot be improved. Notice also that in case $\alpha$ has degree 2, Lemma 4.3.2 gives a better result. An outline of the proof of Roth’s theorem can be found in [Sch91]; complete versions of the proof can be found in [Sch80], [EE93], and [Rot55].

In other words, Roth’s theorem implies that if $\alpha$ is algebraic, then the number of sufficiently good rational approximations to $\alpha$ is finite, so perhaps one can actually count them, although we are not quite ready to do this. If $\alpha$ is real, but not necessarily algebraic, there may be infinitely many good rational approximations to $\alpha$, however we will now show that there are only finitely many of them within a finite interval. To prove a result of this sort, we will first need a certain “gap principle”.

**Definition 4.5.1.** A set $S \subseteq \mathbb{R}$ is called a $C$-set for a real number $C > 1$ if for any two numbers $m, n$ in $S$, $m \leq Cn$ and $n \leq Cm$.

Notice for instance that a $C$-set consisting of integers must be finite, although unless we know at least one of its elements, we cannot say anything about its cardinality.
**Definition 4.5.2.** A set $S \subseteq \mathbb{R}$ is called a $\gamma$-set for a real number $\gamma > 1$ if whenever $m, n \in S$ and $m < n$, then $\gamma m \leq n$.

Notice that a $\gamma$-set can be infinite, but it has a gap principle: its elements cannot be too close together, i.e., there is always a gap between them. A set $S \subseteq \mathbb{Z}_{>0}$ that is both a $C$-set and a $\gamma$-set will be called a $(C, \gamma)$-set. Notice that a $(C, \gamma)$-set is always finite. It is possible to estimate the cardinality of a $(C, \gamma)$-set without knowing anything about its elements.

**Lemma 4.5.2.** Let $C > 1$ and $\gamma > 1$, and suppose that $S \subseteq \mathbb{R}_{>0}$ is a $(C, \gamma)$-set. Then

$$|S| \leq 1 + \frac{\ln C}{\ln \gamma}. \tag{4.13}$$

**Proof.** Clearly $S$ is a finite set, so assume

$$S = \{m_0 < m_1 < \cdots < m_k\},$$

i.e. $|S| = k + 1$. Then for each $0 \leq i \leq k$,

$$m_i \geq m_0 \gamma^i,$$

and

$$Cm_0 \geq m_k \geq m_0 \gamma^k.$$ 

Hence

$$k \leq \frac{\ln C}{\ln \gamma},$$

and (4.13) follows. \hfill \Box

**Definition 4.5.3.** Given $C > 1$, a window of exponential width $C$ is an interval of real numbers $x$ of type

$$w \leq x < w^C,$$

for some $w > 1$.

We can now use Lemma 4.5.2 to prove a bound on the number of good rational approximations to a real number $\alpha$ in a window of exponential width $C$ for any $C > 1$. We will say that a rational number $\frac{p}{q}$ is reduced if $\gcd(p, q) = 1$.

**Lemma 4.5.3.** Let $\alpha \in \mathbb{R}$, $\delta > 0$, and $C > 1$. Let $N_C(\alpha)$ be the number of reduced rational numbers $\frac{p}{q}$ such that

$$|\alpha - \frac{p}{q}| < \frac{1}{2q^2 + \delta} \tag{4.14}$$

and $q$ is in a window of exponential width $C$. Then

$$N_C(\alpha) \leq 1 + \frac{\ln C}{\ln(1 + \delta)} \tag{4.15}.$$

**Proof.** Notice that if $x, y$ are in a window of exponential width $C$, then

$$w \leq x < w^C \leq x^C, \quad w \leq y < w^C \leq y^C,$$
for some \( w > 1 \), hence \( x \leq y^C \) and \( y \leq x^C \). Now suppose that \( \frac{p_1}{q_1} \neq \frac{p_2}{q_2} \) are reduced fractions that satisfy (4.14) with \( 1 \leq q_1 \leq q_2 \) in a window of exponential width \( C \). Then

\[
\frac{1}{q_1 q_2} \leq \left| \frac{p_1}{q_1} - \frac{p_2}{q_2} \right| = \left| \frac{p_1}{q_1} - \alpha \right| + \left| \alpha - \frac{p_2}{q_2} \right| < \frac{1}{2q_1^{2+\delta}} + \frac{1}{2q_2^{2+\delta}} \leq \frac{1}{q_1^{2+\delta}},
\]

and so

\[ q_2 > q_1^{1+\delta}. \]

In other words, if \( q_1 \leq q_2 \) are denominators of the rational approximations \( \frac{p_1}{q_1}, \frac{p_2}{q_2} \) satisfying the hypotheses of the lemma, then

\[ \gamma \ln q_1 < \ln q_2, \]

where \( \gamma = 1 + \delta \), i.e. logarithms of these denominators form a \( \gamma \)-set. On the other hand, if \( q_1, q_2 \) are in a window of exponential width \( C \), then

\[ \ln q_1 \leq C \ln q_2, \quad \ln q_2 \leq C \ln q_1, \]

that is these logarithms also form a \( C \)-set, hence they form a \( (C, \gamma) \)-set, and by Lemma 4.5.2 the cardinality of this set is

\[ \leq 1 + \frac{\ln C}{\ln \gamma} = 1 + \frac{\ln C}{\ln(1 + \delta)}, \]

but this is precisely the number \( N_C(\alpha) \). This completes the proof. \( \square \)

**Remark 4.5.2.** Suppose that \( 1 < A < B \) are given, and suppose that we want to know the number of reduced rational approximations \( \frac{p}{q} \) to the real number \( \alpha \) with

\[ \left| \alpha - \frac{p}{q} \right| < \frac{1}{2q^{2+\delta}}, \]

and \( A \leq q \leq B \). Notice that denominators \( q \) lie in a window of exponential width \( C = \frac{\ln B}{\ln A} \), since

\[ A = e^{\ln A} \leq q \leq B = (e^{\ln A})^{\frac{\ln B}{\ln A}}, \]

and so by Lemma 4.5.3, the number of such approximations is

\[ \leq 1 + \frac{\ln \left( \frac{\ln B}{\ln A} \right)}{\ln(1 + \delta)}. \]

**Definition 4.5.4.** Let \( \alpha \in \mathbb{R} \) and let \( \delta > 0 \). We will call \( \frac{p}{q} \in \mathbb{Q} \) a \( \delta \)-approximation to \( \alpha \) if \( q > 0 \), \( \gcd(p, q) = 1 \), and

\[ \left| \alpha - \frac{p}{q} \right| < \frac{1}{q^{2+\delta}}. \]

A method similar to the proof of Lemma 4.5.3 yields the following result; a proof of this can be found on p. 59 of [Sch91].

**Lemma 4.5.4.** Let \( \alpha \in \mathbb{R}, \delta > 0 \). The number of \( \delta \)-approximations \( \frac{p}{q} \) to \( \alpha \) in a window \( w \leq q \leq w^C \), where \( w \geq 4^{1/\delta} \) is

\[ \leq 1 + \frac{\ln 2C}{\ln(1 + \delta)}. \]
4.6. Continued fractions

In the previous sections we learned about existence and limitations of good rational approximations to an irrational number, however we have not really discussed how to construct such approximations. In this section we introduce a new way of thinking about real numbers, which will yield such a construction.

**Definition 4.6.1.** For a real number \( \alpha \), an expression of the form

\[
\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \ddots}}}}
\]

where \( a_0 \) is an integer and \( a_1, a_2, \ldots \) are positive integers is called a *continued fraction* expansion for \( \alpha \). We will use more compact notation \( \alpha = [a_0; a_1, a_2, a_3, \ldots] \) for this expansion, and for each \( n \geq 1 \) will define its *n-th convergent* to be

\[
[a_0; a_1, a_2, a_3, \ldots, a_n] := a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}
\]

We will call a continued fraction expansion for \( \alpha \) finite if there exists some \( n \) for which \( \alpha \) is equal to its \( n \)-th convergent, and infinite otherwise.

**Theorem 4.6.1.** For each \( \alpha \in \mathbb{R} \) there exists a continued fraction expansion, which is finite if and only if \( \alpha \) is rational.

**Proof.** To prove this result, we present an actual algorithm to compute a continued fraction expansion for \( \alpha \). We do it recursively. Define \( a_0 = \lfloor \alpha \rfloor \), the integer part of \( \alpha \). If \( \alpha = \lfloor \alpha \rfloor \), we are done. If not, then

\[
\alpha = \lfloor \alpha \rfloor + (\alpha - \lfloor \alpha \rfloor) = a_0 + \cfrac{1}{r_1},
\]

where \( r_1 = \frac{1}{\alpha - \lfloor \alpha \rfloor} \). If \( r_1 \) is an integer, we are done. If not, let \( a_1 = \lfloor r_1 \rfloor \), and so

\[
\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{r_2}},
\]

where \( r_2 = \frac{1}{r_1 - \lfloor r_1 \rfloor} \). Continuing in the same manner, on \( n \)-th step we let \( a_{n-1} = \lfloor r_{n-1} \rfloor \), \( r_n = \frac{1}{r_{n-1} - \lfloor r_{n-1} \rfloor} \) and obtain

\[
\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}
\]

This algorithm terminates when \( r_n \) is an integer. If this happens for some \( n \), then bringing all the fractions to a common denominator, we can obtain a rational
expression for $\alpha$, meaning that $\alpha \in \mathbb{Q}$. Hence irrational numbers must have an infinite continued fraction expansion.

On the other hand, suppose $\alpha$ is rational, then each $r_n$ is a rational number, say $r_n = p_n/q_n$, and

$$r_{n+1} = \frac{p_{n+1}}{q_{n+1}} = \frac{1}{r_n - \lfloor r_n \rfloor} = \frac{q_n}{p_n - q_n[p_n/q_n]},$$

and so $p_{n+1} = q_n > q_{n+1}$. This means that the sequence of denominators of $r_1, r_2, \ldots$, namely $q_1, q_2, \ldots$ is decreasing while consisting of positive integers. Hence the algorithm must terminate, i.e. reach the point where some $q_k = 1$ and hence the corresponding $r_k$ is an integer. □

Let us consider some examples. For instance, $7/3 = 2 + \frac{1}{3} = [2; 3]$, as well as

$$-\frac{93}{37} = -3 + \frac{1}{2 + \frac{1}{18}} = [-3; 2, 18], \quad \frac{103}{1647} = 0 + \frac{1}{15 + \frac{1}{102}} = [0; 15, 1]$$

are rational numbers, hence finite continued fractions. On the other hand,

$$\sqrt{2} = \sqrt{2} \left(1 + \sqrt{2} \right) = 1 + \frac{1}{1 + \sqrt{2}} = 1 + \frac{1}{1 + \left(1 + \frac{1}{1 + \sqrt{2}} \right)}$$

and so

$$\sqrt{2} = 1 + \frac{1}{2 + \frac{1}{1 + \left(1 + \frac{1}{1 + \sqrt{2}} \right)}} = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \ldots}}}}}$$

are examples of infinite continued fraction expansions for irrational numbers, which, using our compact notation can also be written as $\sqrt{2} = [1; 2, 2, 2, \ldots]$ and $\pi = [3; 7, 15, 1, 292, 1, 1, 2, 1, 3, \ldots]$.

**Theorem 4.6.2.** For every irrational $\alpha \in \mathbb{R}$ there is a unique continued fraction expansion. If $\alpha$ is rational, there are two continued fraction expansions:

$$\alpha = [a_0; a_1, \ldots, a_n] = [a_0; a_1, \ldots, a_n - 1, 1].$$

**Proof.** First consider the rational case, and notice that indeed

$$a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\ddots + \frac{1}{a_n}}}} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\ddots + \frac{1}{(a_n - 1) + \frac{1}{1}}}}}.$$

Now suppose that $[a_0; a_1, a_2, \ldots]$ is either an infinite continued fraction expansion for a real number $\alpha$ or a finite expansion with the last term $\neq 1$. Let us prove
the uniqueness of this expansion. Suppose that there exists some other continued fraction \([b_0; b_1, b_2, \ldots]\) so that
\[
\alpha = [a_0; a_1, a_2, \ldots] = [b_0; b_1, b_2, \ldots],
\]
then we want to prove that \(a_k = b_k\) for each \(k \geq 0\). We will argue by induction on \(k\). First notice that \(a_0\) must be equal to \(b_0\), since otherwise
\[
|a_0 - b_0| \geq 1 > |[0; a_1, \ldots] - [0; b_1, \ldots]|,
\]
and so we cannot have \([a_0; a_1, \ldots] = [b_0; b_1, \ldots]\). Suppose \(a_k = b_k\) for all \(k \leq m\), then we have
\[
[a_0; a_1, \ldots, a_m, a_{m+1}, \ldots] = [a_0; a_1, \ldots, a_m, b_{m+1}, \ldots],
\]
which implies that there must be equality of the new continued fractions:
\[
[a_{m+1}; a_{m+2}, \ldots] = [b_{m+1}; b_{m+2}, \ldots].
\]
By the same argument as above,
\[
|a_{m+1} - b_{m+1}| \geq 1 > |[0; a_{m+2}, \ldots] - [0; b_{m+2}, \ldots]|,
\]
and so we must have \(a_{m+1} = b_{m+1}\). We should stress that the inequalities in (4.17) and (4.18) are strict: the only other option would be for the continued fraction to be finite with the last term being equal to 1, which we assumed is not the case. This completes the proof by induction. \(\square\)

Since each \(n\)-th convergent \(\alpha_n\) of an irrational number \(\alpha\) is rational, we can write it as \(\alpha_n = \frac{p_n}{q_n}\) for some integers \(p_n\) and \(q_n\). It is now natural to approximate a real number \(\alpha\) by its convergents \(\alpha_n\). How close is this approximation?

**Theorem 4.6.3.** Let \(\alpha \in \mathbb{R}\) and let \(\alpha_n = \frac{p_n}{q_n}\) be its \(n\)-th convergent. Then for every integer \(n \geq 1\),
\[
\left| \alpha - \frac{p_n}{q_n} \right| \leq \frac{1}{q_nq_{n+1}}.
\]

To prove this theorem, we first need a couple of auxiliary lemmas.

**Lemma 4.6.4.** With the notation of Theorem 4.6.3, for every \(n \geq 1\)
\[
p_n = a_nn_{n-1} + p_{n-2}, \quad q_n = a_nq_{n-1} + q_{n-2}.
\]

**Proof.** We argue by induction on \(n\). First notice that \(\frac{p_0}{q_0} = \frac{a_0}{1}\) and \(\frac{p_1}{q_1} = \frac{a_0a_1+1}{a_1}\). If \(n = 2\), we have
\[
\frac{p_2}{q_2} = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = \frac{a_0a_1a_2 + a_2 + a_0}{a_1a_2 + 1} = \frac{a_2p_1 + p_0}{a_2q_1 + q_0}.
\]
Now assume the statement is true for all \(n \leq k - 1 \geq 2\), and let us prove it for \(n = k\). Let us define
\[
t_{k-1} = \frac{a_1; a_2, \ldots, a_{k-1}}{s_{k-1}}
\]
to be the \((k-1)\)-st convergent of the derived continued fraction \([a_1; a_2, \ldots]\). Then induction hypothesis applies to \(t_{k-1}/s_{k-1}\), and
\[
\frac{p_k}{q_k} = a_0 + \frac{s_{k-1}}{t_{k-1}} = \frac{a_0t_{k-1} + s_{k-1}}{t_{k-1}} = \frac{a_0(a_kt_{k-1} + t_{k-2}) + a_ks_{k-1} + s_{k-2}}{a_kt_{k-1} + t_{k-2}}
\]
\[
= \frac{a_k(a_0t_{k-1} + s_{k-1}) + (a_0t_{k-2} + s_{k-2})}{a_kt_{k-1} + t_{k-2}}.
\]
Now, Problem 4.10 guarantees that for each $n$,

$$p_n = a_0 t_n + s_n, \quad q_n = t_n.$$  

Substituting this into the above equation with $n = k - 1$ and $k - 2$, we obtain

$$\frac{p_k}{q_k} = \frac{a_k p_{k-1} + p_{k-2}}{a_k q_{k-1} + q_{k-2}}.$$  

This completes the proof of this lemma. \qed

**Lemma 4.6.5.** With the notation of Theorem 4.6.3, for every $n \geq 1$

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

**Proof.** Multiplying both sides of the above equation by $q_{n-1} q_n$, we see that we need to prove

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

Again, we argue by induction on $n$. If $n = 1$, then

$$p_0 = a_0, \quad q_0 = 1, \quad p_1 = a_0 a_1 + 1, \quad q_1 = a_1,$$

and so

$$p_0 q_1 - p_1 q_0 = a_0 a_1 - (a_0 a_1 + 1) = -1 = (-1)^1.$$  

Assume now that the statement is proved for $n \leq k - 1$ and let us prove it for $n = k$. Applying Lemma 4.6.4 along with the induction hypothesis, we have:

$$\begin{align*}
p_{k-1} q_n - p_k q_{k-1} & = p_{k-1} (a_k q_{k-1} + q_{k-2}) - (a_k p_{k-1} + p_{k-2}) q_{k-1} \\
& = p_{k-1} q_{k-2} - p_{k-2} q_{k-1} = (-1)^{k-1} = (-1)^k.
\end{align*}$$

\qed

**Corollary 4.6.6.** With the notation of Theorem 4.6.3,

$$\frac{p_{2k}}{q_{2k}} \leq \alpha, \quad \frac{p_{2k+1}}{q_{2k+1}} \geq \alpha$$

for all $k \geq 0$.

**Proof.** By Lemma 4.6.5,

$$\frac{p_{n-2}}{q_{n-2}} - \frac{p_n}{q_n} = \left( \frac{p_{n-2}}{q_{n-2}} - \frac{p_{n-1}}{q_{n-1}} \right) + \left( \frac{p_{n-1}}{q_{n-1}} - \frac{p_n}{q_n} \right)$$

$$= (-1)^{n-1} + (-1)^n = (-1)^{n-1} \left( \frac{1}{q_{n-2}} - \frac{1}{q_n} \right),$$

which is positive for odd $n$ and negative for even, since the sequence of denominators $q_n$ is increasing by Lemma 4.6.4. Therefore, when $n$ is even the sequence of convergents $p_n/q_n$ is increasing, and when $n$ is odd it is decreasing. Since in both cases the convergents tend to $\alpha$, the conclusion follows. \qed

**Proof of Theorem 4.6.3.** We can now prove the theorem. By Corollary 4.6.6, the odd-numbered convergents are greater or equal than $\alpha$ and the even-numbered ones are less or equal than $\alpha$. Therefore

$$\frac{|p_n - p_{n+1}|}{q_n} = \left| \frac{p_n}{q_n} - \alpha \right| + \left| \frac{\alpha - p_{n+1}}{q_{n+1}} \right| = \left| \frac{p_n}{q_n} \right| + \left| \frac{\alpha - p_{n+1}}{q_{n+1}} \right|.$$
On the other hand, Lemma 4.6.5 guarantees that
\[
\left| \frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}} \right| = \frac{1}{q_n q_{n+1}}.
\]
Combining these two observations yields the result. \(\Box\)

**Theorem 4.6.7.** Let \(\alpha \in \mathbb{R}\) be irrational, and let \(\alpha_n = \frac{p_n}{q_n}\) be its \(n\)-th convergent. Then for any rational number \(p/q\) with \(q \leq q_n\),
\[
\left| \alpha - \frac{p_n}{q_n} \right| \leq \left| \alpha - \frac{p}{q} \right|.
\]
In other words, \(p_n/q_n\) is the best rational approximation to \(\alpha\) among all rational numbers with denominators no bigger than \(q_n\).

**Proof.** Let \(p/q \neq p_n/q_n\) be any rational approximation to \(\alpha\) with \(q \leq q_n\). Let
\[
A = \begin{pmatrix} p_n & p_{n+1} \\ q_n & q_{n+1} \end{pmatrix},
\]
then
\[
\det(A) = p_n q_{n+1} - p_{n+1} q_n = (-1)^{n+1}
\]
by Lemma 4.6.5. Therefore the lattice \(A \mathbb{Z}^2 = \mathbb{Z}^2\), so there exist \(x, y \in \mathbb{Z}\) such that
\[
A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix},
\]
in other words
\[(4.19) \quad p = x p_n + y p_{n+1}, \quad q = x q_n + y q_{n+1}.
\]
We cannot have \(x\) and \(y\) both equal 0. Suppose that \(x = 0, y \neq 0\), then \(q = y q_{n+1} > q_n\) by Lemma 4.6.4, which contradicts the choice of \(q\). Then assume \(y = 0, x \neq 0\), then \(q = x q_n\), so \(x = 1\), and hence \(p = p_n, q = q_n\), again a contradiction. Hence we must have \(x, y \neq 0\). Further,
\[
0 < q = x q_n + y q_{n+1} \leq q_n < q_{n+1},
\]
thus \(x\) and \(y\) must have different signs. Notice that
\[
q_n \alpha - p_n \quad \text{and} \quad q_{n+1} \alpha - p_{n+1}
\]
must also have different signs, by Corollary 4.6.6. Therefore
\[
|q \alpha - p| = |x(q_n \alpha - p_n)| + |y(q_{n+1} \alpha - p_{n+1})| > |x(q_n \alpha - p_n)| \geq |q_n \alpha - p_n|.
\]
Then:
\[
\left| \alpha - \frac{p}{q} \right| > \frac{|q \alpha - p|}{q} > \frac{|q_n \alpha - p_n|}{q} = \frac{1}{q} \geq \left| \alpha - \frac{p_n}{q_n} \right|.
\]
This completes the proof. \(\Box\)

We proved that continued fraction expansion can be used to provide best rational approximations to an irrational number. This fact can be interpreted geometrically. Suppose \(\alpha\) is an irrational number, and let \(p_n/q_n\) be its \(n\)-th convergent. Then \((q_n, p_n)\) is an integer lattice point in the plane: it is the closest lattice point to the line \(y = \alpha x\) in the box
\[
\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq q_n, 0 \leq y \leq p_n\}.\]
In this way, the lines with rational slope \( y = \frac{p}{q} x \) approximate the line \( y = \alpha x \) with irrational slope, and no line through the origin and a lattice point inside this box can come closer. Let us consider an example. Let
\[
\alpha = \sqrt{2} = [1; 2, 2, 2, \ldots] = 1.41421356237 \ldots,
\]
then:
\[
\alpha_1 = \frac{p_1}{q_1} = [1; 2] = 1 + \frac{1}{2} = \frac{3}{2} = 1.5
\]
\[
\alpha_2 = \frac{p_2}{q_2} = [1; 2, 2] = 1 + \frac{1}{2 + \frac{1}{2}} = \frac{7}{5} = 1.4
\]
\[
\alpha_3 = \frac{p_3}{q_3} = [1; 2, 2, 2] = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}} = \frac{17}{12} = 1.4166666666\ldots
\]
\[
\alpha_4 = \frac{p_4}{q_4} = [1; 2, 2, 2, 2] = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}}} = \frac{41}{29} = 1.41379310345
\]

**Figure 1.** Continued fraction convergent approximations to \( \sqrt{2} \)

These are the best rational approximations to \( \sqrt{2} \), where each odd numbered convergent yields a lattice point above the line \( y = \sqrt{2}x \) and each even numbered one provides a point below this line. For a detailed exposition of the theory of continued fractions see [Kar13].
4.7. Kronecker’s theorem

Here we briefly mention a famous theorem of Leopold Kronecker, which is obtained by an application of Dirichlet’s approximation theorem. Let \( \{ \alpha \} \) denote the fractional part of the real number \( \alpha \).

**Theorem 4.7.1 (Kronecker).** The sequence of fractional parts \( \{ n\alpha \} \) as \( n \) ranges over all positive integers is dense in the interval \([0, 1]\) if and only if \( \alpha \in \mathbb{R} \) is irrational.

**Proof.** Problem 4.17 asserts that the sequence \( \{ n\alpha \} \) is periodic with a finite period for rational \( \alpha \), hence it cannot be dense. We now prove that in case \( \alpha \) is irrational the sequence is indeed dense in \([0, 1]\). For a real number \( b \), let us write \( \| b \| \) for its distance to the nearest integer. For instance, \( \| 3.14 \| = 0.14 \) while \( \| 2.71 \| = 0.29 \). Let \( x \in [0, 1] \) and let \( \varepsilon > 0 \). It is sufficient to show that there exists some positive integer \( n \) so that

\[
\| n\alpha - x \| < \varepsilon.
\]

By Dirichlet’s Theorem 4.3.1, there exist infinitely many rationals \( p/q \) with \( \gcd(p, q) = 1 \) such that \( |\alpha - p/q| < \frac{1}{q^2} \). Let us take \( q > 1/\varepsilon \), then we have

\[
0 < \| \alpha \| \leq |\alpha q - p| < \frac{1}{q} < \varepsilon.
\]

This inequality implies that either

\[
0 < \alpha q - p < 1/q
\]

or

\[
-1/q < \alpha q - p < 0.
\]

First assume (4.21) holds. Subdivide the interval \([0, 1]\) into subintervals of length \( \alpha q - p \) (the last one will be shorter). Then \( x \) falls into one of these intervals, say into the \( m \)-th one for some integer \( m \geq 1 \). If this is not the last such subinterval, then \( m(\alpha q - p) \) is the right end-point of it; if it is the last one, then \( (m-1)(\alpha q - p) \) is its left end-point. In any case, we have

\[
\| \ell q\alpha - x \| < \| \ell(\alpha q - p) - x \| < |\alpha q - p| < \frac{1}{q} < \varepsilon,
\]

where \( \ell = m \) or \( m - 1 \). Then (4.20) holds with \( n = \ell q \). The argument is the same if (4.22) holds instead, where we simply replace \( \alpha q - p \) with \( p - \alpha q \). \( \square \)

**Remark 4.7.1.** More generally, it is also true that for any integer \( k \geq 1 \) the sequence of fractional parts \( \{ n^k \alpha \} \) is dense in the interval \([0, 1]\) if and only if \( \alpha \in \mathbb{R} \) is irrational. Further, for irrational \( \alpha \) the sequence \( \{ n^k \alpha \} \) is equidistributed in the interval \([0, 1]\) for every \( k \geq 1 \): a sequence \( \{ x_n \}_{n=1}^{\infty} \) is said to be equidistributed in the interval \([0, 1]\) if

\[
\lim_{T \to \infty} \frac{|\{ n : n \leq T, x_n \in [a, b] \}|}{T} = b - a
\]

for any \( 0 \leq a < b \leq 1 \). We refer the reader to [MTB06], Chapter 12 for some details.
The general version of Kronecker’s theorem gives more. Suppose $1, \alpha_1, \ldots, \alpha_m$ are real numbers, which are linearly independent over $\mathbb{Q}$. Then the sequence of points

$$\left(\{n\alpha_1\}, \ldots, \{n\alpha_m\}\right)_{n=1}^{\infty}$$

is dense in the unit cube $[0, 1]^m \subset \mathbb{R}^m$. A detailed account of this multi-dimensional theorem, as well as a general theory of simultaneous Diophantine approximation can be found in Cassels’ classical book [Cas57]. A survey of some more recent results in the direction of Kronecker’s theorem is given in [GM16]; see also [FM18] for a very general effective version of this theorem.

We conclude this chapter by another remarkable result related to Kronecker’s theorem, known as the Three-Gap Theorem.

**Theorem 4.7.2 (Three-Gap Theorem).** Suppose that $n$ points have been placed on a circle at angles $\theta, 2\theta, \ldots, n\theta$ from the starting point. Then there can be at most three distinct distances between adjacent pairs of these points around the circle.

This observation was first conjectured by Hugo Steinhaus, and then proved in the 1950’s by Vera Sós, János Surányi, and Stanisław Świerczkowski. The elegant proof we discuss here is very recent: it is due to Marklof and Strömbergsson [MS17], and is based on the geometry of lattices. We outline only a brief sketch of their argument. Let us think of the angles as parts of the circle with the full angle $2\pi$ being 1. Hence all the angular positions can be thought as real numbers in the interval $[0, 1]$, where the endpoints have been identified: this is precisely a set of coset representatives of the quotient additive group $\mathbb{R}/\mathbb{Z}$. Let $\alpha$ be the angular position of $\theta$, then angular positions of the angles $\theta, 2\theta, \ldots, n\theta$ are given by the sequence of fractional parts

$$\left(\xi_k\right)_{k=1}^{n} = \{k\alpha\}.$$ 

Then the distances between our angular positions on the circle are precisely the gaps between corresponding numbers in this sequence $\left(\xi_k\right)_{k=1}^{n}$. The gap between $\xi_k$ and its next neighbor in $\mathbb{R}/\mathbb{Z}$ (in the direction to the right, so this is not necessarily nearest neighbor, as the nearest may be on the left) is

$$s_{k,n} = \min \left\{ (\ell - k)\alpha + n > 0 : (\ell, n) \in \mathbb{Z}^2, 0 < \ell \leq n \right\} = \min \left\{ m\alpha + n > 0 : (m, n) \in \mathbb{Z}^2, -k < m \leq n - k \right\},$$

where the second equality is obtained by the substitution $m = \ell - k$. Let

$$A_1 = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix},$$

and notice that

$$s_{k,n} = \min \left\{ y > 0 : (x, y)^\top \in A_1 \mathbb{Z}^2, -k < x \leq n - k \right\}.$$ 

This way $s_{k,n}$ can be thought of as a function on the lattice $A_1 \mathbb{Z}^2$. More generally, we can define

$$F(M, t) = \min \left\{ y > 0 : (x, y)^\top \in M \mathbb{Z}^2, -t < x \leq 1 - t \right\}.$$ 

This is a function $F : \text{SL}_2(\mathbb{R}) \times (0, 1] \to \mathbb{R}_{>0}$, where $\text{SL}_2(\mathbb{R}) = \{M \in \text{GL}_2(\mathbb{R}) : \det(M) = 1\}$ is a subgroup of $\text{GL}_2(\mathbb{R})$ consisting of matrices with unit determinant.
In fact, as the authors show in [MS17] (see Proposition 1), $F(M, t)$ is well-defined as a function on the space of lattices for every fixed $t$, i.e. $F(M, t) = F(M', t)$ if $M$ and $M'$ are two basis matrices for the same lattice. Define

$$A_n = \begin{pmatrix} n^{-1} & 0 \\ 0 & n \end{pmatrix} A_1 \in \text{SL}_2(\mathbb{R}),$$

and notice that

$$s_{k,n} = \frac{1}{n} \min \left\{ y > 0 : (x, y)^\top \in A_n \mathbb{Z}^2, -\frac{k}{n} < x \leq 1 - \frac{k}{n} \right\} = \frac{1}{n} F(A_n, k/n).$$

Hence the proof of Theorem 4.7.2 is reduced to showing that for every $M \in \text{SL}_2(\mathbb{R})$, the function $t \to F(M, t)$ is piecewise constant and takes on at most three distinct values; in fact, if there are three values, then the third is the sum of the first and second. This is the assertion of Proposition 2 of [MS17].
4.8. Problems

Problem 4.1. Let \( \{a_k\}_{k=0}^{\infty} \) be a sequence of integers in some interval \([b, c]\), where \( b < c \) are constants. Prove that the power series
\[
\sum_{k=0}^{\infty} a_k 10^{-k},
\]
is absolutely convergent.

Problem 4.2. Let \( \alpha \in A \) and let \( f(x), g(x) \in \mathbb{Z}[x] \) be two polynomials of degree \( \deg(\alpha) \) such that \( f(\alpha) = g(\alpha) = 0 \). Prove that \( f(x) = cg(x) \) for some constant \( c \).

Problem 4.3. Prove that \( m_{\alpha}(x) \) is irreducible for each \( \alpha \in A \). Furthermore, prove that if \( p(x) \in \mathbb{Z}[x] \) is such that \( p(\alpha) = 0 \), then \( m_{\alpha}(x) \mid p(x) \), i.e. there exists some \( g(x) \in \mathbb{Z}[x] \) such that \( p(x) = m_{\alpha}(x)g(x) \).

Problem 4.4. Prove that any infinite subset of a countable set is countable. Use this fact to conclude that a superset of an uncountable set is uncountable.

Problem 4.5. Let \( a < b \) be real numbers and let \( I = [a, b] \) be a closed interval. Prove that \( I \) contains infinitely many real numbers.

Problem 4.6. A subset \( S \) of \( \mathbb{R} \) is called discrete if there exists real \( \varepsilon > 0 \) such that for every two distinct elements \( \alpha, \beta \in S \),
\[
|\alpha - \beta| \geq \varepsilon.
\]
On the other hand, let us say that \( S \) is near-discrete if for every \( \alpha \in S \) there exist a real \( \varepsilon = \varepsilon(\alpha) > 0 \) such that
\[
|\alpha - \beta| \geq \varepsilon
\]
for every \( \beta \in S \) distinct from \( \alpha \).

a) Prove that every discrete subset of \( \mathbb{R} \) is countable.
b) Prove that every near-discrete subset of \( \mathbb{R} \) is countable.

Problem 4.7. Prove that if \( \alpha = \frac{a}{Q+1} \) for some integer \( a \) with
\[
\gcd(a, Q + 1) = 1,
\]
then there is equality in (4.2).

Problem 4.8. Let \( \mu > 0 \). We say that \( p/q \in \mathbb{Q} \) is a \( \mu \)-approximation to the real number \( \alpha \) if
\[
\left| \frac{\alpha - p}{q} \right| < \frac{1}{q^\mu}.
\]
Let \( S \subseteq \mathbb{R} \) be a set with the following properties:

(1) Every element of \( S \) has infinitely many rational 3-approximations.
(2) If a rational number \( p/q \) is a 3-approximation for some \( \alpha \in S \), then it is not a 3-approximation for any other element of \( S \).
a) Prove that $S$ is countable.
b) Prove that every $\alpha \in S$ is transcendental.

**Problem 4.9.** Let $\{n_j\}_{j=1}^\infty$ be a sequence of natural numbers such that
\[
\lim_{j \to \infty} \frac{n_{j+1}}{n_j} = \infty.
\]
Define
\[
f(x) = \sum_{j=1}^\infty x^{n_j},
\]
and let $\alpha$ be a real algebraic number such that $0 < \alpha < 1$. Prove that $f(\alpha)$ is transcendental.

**Problem 4.10.** Let $\alpha$ have the continued fraction expansion $[a_0; a_1, a_2, \ldots]$, and let $p_n/q_n$ be its $n$-th convergent. Let $t_n/s_n$ be the $n$-th convergent of the number $[a_1; a_2, \ldots]$. Prove that
\[
p_n = a_0t_n + s_n, \quad q_n = t_n.
\]

**Problem 4.11.** The golden ratio is defined as
\[
\phi = \frac{1 + \sqrt{5}}{2}.
\]
This number appears in numerous places in mathematics, architecture, engineering and science throughout history, starting with proportion calculations for particularly symmetric structures in ancient Egypt (e.g. Great Pyramid of Giza) and then ancient Greece and Rome. Prove that
\[
\phi^2 = \phi + 1,
\]
and derive from it that
\[
\phi = 1 + \frac{1}{\phi}.
\]

**Problem 4.12.** Derive a continued fraction expansion for $\phi$ $[a_0; a_1, a_2, \ldots]$ and use it to prove that $\phi$ is irrational.

**Problem 4.13.** Let us write $\phi_n$ for the $n$-th continued fraction approximations to $\phi$, i.e.
\[
\phi_n = [a_0; a_1, \ldots, a_n].
\]
Compute $\phi_1$ through $\phi_5$ as fractions.
Problem 4.14. Now let us define the Fibonacci sequence. Let

\[ F_1 = 1, F_2 = 1, \]

and for every \( n \geq 3 \), let

\[ F_n = F_{n-1} + F_{n-2}. \]  

This sequence is named after the 12th century Italian mathematician Leonardo Fibonacci of Pisa, but its origins go back to the study of poetic structures in Sanskrit in ancient India. These numbers are so important in mathematics, science and engineering that there are many things named after them, including the mathematical journal Fibonacci Quarterly devoted entirely to the study of the Fibonacci sequence and its many connections.

Compute the first ten Fibonacci numbers.

Problem 4.15. Now compute as fractions the ratios of the consecutive Fibonacci numbers

\[
\frac{F_3}{F_2}, \frac{F_4}{F_3}, \frac{F_5}{F_4}, \frac{F_6}{F_5}, \frac{F_7}{F_6}.
\]

Compare with convergents of the golden ratio.

Problem 4.16. Prove the general formula:

\[ \phi_n = \frac{F_{n+2}}{F_{n+1}}, \]

for every \( n \geq 1 \).

Problem 4.17. Let \( \alpha = p/q \) with \( \gcd(p,q) = 1 \). Prove that the sequence of fractional parts \( \{n\alpha\} \) as \( n \) ranges over positive integers is periodic with period \( q \).

Problem 4.18. Prove that the sequence \( a_n = \sin n \) as \( n \) ranges over all the integers is dense in the interval \([-1, 1]\).

Hint: To prove that \( \sin n \) comes arbitrary close to any \( \beta \in [-1, 1] \), let \( \alpha \in [0, 1) \) be such that \( \beta = \sin(2\pi\alpha) \), apply Kronecker’s theorem and use continuity of \( \sin x \).
CHAPTER 5

Algebraic Number Theory

5.1. Some field theory

Our next goal is to develop some further properties of algebraic and transcendental numbers. For this we need to introduce some elements of field theory.

Definition 5.1.1. Let $K$ and $L$ be fields with the same addition and multiplication operations such that $K \subseteq L$. Then $L$ is called a field extension of $K$, denoted $L/K$, and $K$ is called a subfield of $L$.

If $L$ is a field extension of $K$, then $L$ is a $K$-vector space (Problem 5.1). Its dimension is called the degree of this field extension, denoted by $[L : K]$. If the degree is finite, we say that $L/K$ is a finite extension. A classical example of field extensions comes from extending a subfield of $\mathbb{C}$ (often $\mathbb{Q}$) by some collection of complex numbers. Let $K \subseteq \mathbb{C}$ be a subfield, $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, and define $K(\alpha_1, \ldots, \alpha_n)$ to be the smallest subfield of $\mathbb{C}$ with respect to inclusion that contains $K$ and $\alpha_1, \ldots, \alpha_n$. A subfield $K$ of $\mathbb{C}$ is called algebraic if every element $\alpha \in K$ is an algebraic number.

We will also say that $L/K$ is an algebraic extension if $K \subseteq L \subseteq \mathbb{C}$ are algebraic fields.

Definition 5.1.2. Let $K \subseteq \mathbb{C}$ and $\alpha \in \mathbb{C}$. We define

$$K[\alpha] := \text{span}_{K} \{1, \alpha, \alpha^2, \ldots\}$$

$$= \left\{ \sum_{m=0}^{n} a_m \alpha^m : a_0, \ldots, a_n \in K, \ n \in \mathbb{Z}_{\geq 0} \right\},$$

i.e., the set of all finite linear combinations of powers of $\alpha$ with coefficients from $K$ (notice that this notation is consistent with the notation $K[x]$ denoting the ring of one-variable polynomials with coefficients in $K$ (Problem 5.7)). Then $K[\alpha]$ is a vector space over $K$, whose dimension $\dim_K K[\alpha]$ is equal to the number of powers of $\alpha$ which are linearly independent over $K$. In fact, $K[\alpha] \subseteq K(\alpha)$ (Problem 5.5).

We now establish some important properties of algebraic numbers.

Theorem 5.1.1. Let $\alpha \in \mathbb{C}$.

1. If $\alpha$ is transcendental, then $\dim_\mathbb{Q} \mathbb{Q}[\alpha] = \infty$.
2. If $\alpha$ is algebraic of degree $n$, then $\mathbb{Q}[\alpha] = \text{span}_\mathbb{Q} \{1, \alpha, \ldots, \alpha^{n-1}\}$, and $1, \alpha, \ldots, \alpha^{n-1}$ are linearly independent over $\mathbb{Q}$. Hence

$$\dim_\mathbb{Q} \mathbb{Q}[\alpha] = n.$$

3. $\mathbb{Q}[\alpha]$ is a field if and only if $\alpha$ is algebraic.
4. If $\alpha$ is algebraic, then $\mathbb{Q}(\alpha) = \mathbb{Q}[\alpha]$. 

82
Proof. To prove part (1), assume that \( \dim \mathbb{Q}[\alpha] = n < \infty \). Then the collection of \( n + 1 \) elements \( 1, \alpha, \ldots, \alpha^n \) must be linearly dependent, i.e. there exist \( c_0, \ldots, c_n \in \mathbb{Q} \) such that
\[
c_0 + c_1 \alpha + \cdots + c_n \alpha^n = 0.
\]
Clearing the denominators, if necessary, we can assume that \( c_0, \ldots, c_n \in \mathbb{Z} \), and hence \( \alpha \) is a root of \( \sum_{m=0}^{n} c_m x^m \in \mathbb{Z}[x] \), which means that it is algebraic.

To prove part (2), assume that \( \alpha \) is algebraic of degree \( n \) and let
\[
m_\alpha(x) = \sum_{m=0}^{n} a_m x^m \in \mathbb{Z}[x],
\]
where \( a_n \neq 0 \) and \( a_0 \neq 0 \), since \( m_\alpha(x) \) is irreducible. Since \( m_\alpha(\alpha) = 0 \), we have
\[
\alpha^n = \sum_{m=0}^{n-1} \left( -\frac{a_m}{a_n} \right) \alpha^m.
\]
Therefore any \( \mathbb{Q} \)-linear combination of powers of \( \alpha \) can be expressed as a \( \mathbb{Q} \)-linear combination of \( 1, \alpha, \ldots, \alpha^{n-1} \). Now suppose \( 1, \alpha, \ldots, \alpha^{n-1} \) are linearly dependent, then there exist \( c_0, \ldots, c_{n-1} \in \mathbb{Q} \) such that
\[
c_0 + c_1 \alpha + \cdots + c_{n-1} \alpha^{n-1} = 0.
\]
In fact, clearing the denominators if necessary, we can assume that \( c_0, \ldots, c_{n-1} \in \mathbb{Z} \).

But this means that \( \alpha \) is a root of the polynomial
\[
p(x) = \sum_{m=0}^{n-1} c_m x^m \in \mathbb{Z}[x],
\]
which has degree \( n - 1 \). This contradicts the assumption that \( \deg(\alpha) = n \), hence \( 1, \alpha, \ldots, \alpha^{n-1} \) must be linearly independent, so they form a basis for \( \mathbb{Q}[\alpha] \) over \( \mathbb{Q} \).

For part (3), assume first that \( \alpha \in \mathbb{C} \) is algebraic. It is clear that \( \mathbb{Q}[\alpha] \) is closed under addition and multiplication. We only need to prove that for any \( \beta \in \mathbb{Q}[\alpha] \setminus \{0\} \), there exists \( \beta^{-1} \in \mathbb{Q}[\alpha] \). By part (2), there exist \( b_0, \ldots, b_{n-1} \in \mathbb{Q} \) such that
\[
\beta = \sum_{m=0}^{n-1} b_m \alpha^m.
\]
We want to prove the existence of
\[
\gamma = \sum_{m=0}^{n-1} c_m \alpha^m \in \mathbb{Q}[\alpha]
\]
such that \( \beta \gamma = 1 \). Let \( \gamma \) be as in (5.2) with coefficients \( c_0, \ldots, c_{n-1} \) to be specified, then:
\[
\beta \gamma = \sum_{m=0}^{n-1} \sum_{k=0}^{n-1} b_m c_k \alpha^{m+k} = \sum_{l=0}^{2n-2} \left( \sum_{m+k=l} b_m c_k \right) \alpha^l.
\]
For each \( l \geq n \), we can substitute (5.1) for \( \alpha^n \), lowering the power. After a finite number of such substitutions, we will obtain an expression
\[
\beta \gamma = \sum_{l=0}^{n-1} f_l(c_0, \ldots, c_{n-1}) \alpha^l,
\]
where \( f_l(c_0, \ldots, c_{n-1}) \) is a homogeneous linear polynomial in variables \( c_0, \ldots, c_{n-1} \) with coefficients depending on \( b_i \)’s and \( a_i \)’s, for each \( 0 \leq l \leq n - 1 \). Since we want \( \beta \gamma = 1 \), we set
\[
\begin{align*}
f_0(c_0, \ldots, c_{n-1}) &= 1 \\
f_1(c_0, \ldots, c_{n-1}) &= 0 \\
&\vdots \\
f_{n-1}(c_0, \ldots, c_{n-1}) &= 0.
\end{align*}
\]
(5.4)

This is a linear system of \( n \) equations in \( n \) variables, which can be written as \( Fc = e_1 \), where \( F \) is the \( n \times n \) coefficient matrix of linear polynomials \( f_0, \ldots, f_{n-1} \), \( e_1 = (1, 0, \ldots, 0)^t \in \mathbb{R}^n \) is the first standard basis vector in \( \mathbb{R}^n \), and \( c = (c_0, \ldots, c_{n-1})^t \).

The matrix \( F \) must be nonsingular matrix. Indeed, suppose it is singular, then there exists some \( 0 \neq c \in \mathbb{Q}^n \) such that \( Fc = 0 \), i.e.
\[ f_l(c_0, \ldots, c_{n-1}) = 0 \quad \forall \quad 0 \leq l \leq n - 1. \]

Let \( \gamma \) as in (5.2) be defined with this choice of the coefficient vector \( c \). Then, by (5.3), \( \beta \gamma = 0 \), while \( \beta, \gamma \neq 0 \). This is a contradiction, since there can be no zero divisors in the field \( \mathbb{C} \). Therefore (5.4) has a unique solution \( c \). Let \( \gamma \) be as in (5.2) with this choice of \( c \), then \( \gamma = \beta^{-1} \in \mathbb{Q}[\alpha] \), and so \( \mathbb{Q}[\alpha] \) is a field.

Finally we establish part (4) by proving that \( \mathbb{Q}[\alpha] = \mathbb{Q}(\alpha) \). First notice that \( \mathbb{Q}[\alpha] \subseteq \mathbb{Q}(\alpha) \), since every \( \mathbb{Q} \)-linear combination of powers of \( \alpha \) must be contained in any field containing \( \mathbb{Q} \) and \( \alpha \). To show containment the other way, notice that, by part (3), \( \mathbb{Q}[\alpha] \) is a field containing \( \mathbb{Q} \) and \( \alpha \), and so it must contain \( \mathbb{Q}(\alpha) \). \( \square \)

**Example 5.1.1.** We give an example of finding the inverse of an element of \( \mathbb{Q}[\alpha] \) when \( \alpha \) is algebraic. Consider
\[ \beta = 3^{2/3} + 2 \times 3^{1/3} - 2 \in \mathbb{Q}[3^{1/3}]. \]

We look for
\[ \beta^{-1} = a3^{2/3} + b3^{1/3} - c \in \mathbb{Q}[3^{1/3}]. \]
Then we need
\[
1 = \beta \beta^{-1} = a^{3^{1/3}} + b^{3^{1/3}} - c^{2^{1/3}} + 2a^{3^{1/3}} + 2b^{3^{2/3}} - 2c^{1^{1/3}} \\
-2a^{3^{2/3}} - 2b^{3^{1/3}} + 2c \\
= (2b - 2a - c)^{3^{2/3}} + (3a - 2c - 2b)^{3^{1/3}} + (2c + 6a + 3b),
\]
in other words we are looking for \(a, b, c \in \mathbb{Q}\) such that
\[
2b - 2a - c = 0, \\
3a - 2c - 2b = 0, \\
2c + 6a + 3b = 1.
\]
This system has a unique solution:
\[
a = \frac{6}{61}, b = \frac{7}{61}, c = \frac{2}{61},
\]
hence
\[
\beta^{-1} = \frac{6}{61} 3^{2/3} + \frac{7}{61} 3^{1/3} - \frac{2}{61}.
\]

An immediate consequence of Theorem 5.1.1 is an algebraic criterion for transcendence.

**Corollary 5.1.2.** A number \(\alpha \in \mathbb{C}\) is transcendental if and only if \([\mathbb{Q}(\alpha) : \mathbb{Q}] = \infty\).

**Proof.** If \(\alpha\) is transcendental, then \(\text{dim}_\mathbb{Q} \mathbb{Q}[\alpha] = \infty\) by part (1) of Theorem 5.1.1. On the other hand, \(\mathbb{Q}[\alpha] \subseteq \mathbb{Q}(\alpha)\) by Problem 5.5. Hence \(\mathbb{Q}(\alpha)\) must be an infinite-dimensional \(\mathbb{Q}\)-vector space, hence \([\mathbb{Q}(\alpha) : \mathbb{Q}] = \infty\).

Conversely, suppose that \([\mathbb{Q}(\alpha) : \mathbb{Q}] = \infty\). Assume, towards a contradiction, that \(\alpha\) is algebraic of degree \(n\). By part (4) of Theorem 5.1.1, we have \(\mathbb{Q}(\alpha) = \mathbb{Q}[\alpha]\), but by part (2) of Theorem 5.1.1
\[
\infty \geq n = \text{dim}_\mathbb{Q} \mathbb{Q}[\alpha] = [\mathbb{Q}(\alpha) : \mathbb{Q}].
\]
This is a contradiction, and hence \(\alpha\) must be transcendental. \(\Box\)

Another important consequence is the following.

**Theorem 5.1.3.** The set \(\mathbb{A}\) of algebraic numbers is a field under the usual addition and multiplication of complex numbers.

**Proof.** By Theorem 4.2.5, we know that \(\mathbb{A}\) is countable, and so we can write
\[
\mathbb{A} = \{\alpha_1, \alpha_2, \alpha_3, \ldots\},
\]
choosing an ordering on \(\mathbb{A}\). For each \(n \in \mathbb{N}\), define
\[
K_n := \mathbb{Q}(\alpha_1, \ldots, \alpha_n).
\]
By Problem 5.3, the degree \([K_n : \mathbb{Q}] < \infty\). Let \(n \in \mathbb{N}\) and let \(\beta \in K_n\), then \(\mathbb{Q}(\beta) \subseteq K_n\), which means that
\[
[\mathbb{Q}(\beta) : \mathbb{Q}] \leq [K_n : \mathbb{Q}] < \infty,
\]
and so \(\beta\) is algebraic, by Theorem 5.1.1. Therefore any element of any field \(K_n\) is in \(\mathbb{A}\), and hence we have
\[
\mathbb{Q} \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq \mathbb{A}.
\]
Now let \(0 \neq \beta, \gamma \in \mathbb{A}\), then there exist some integers \(1 \leq k \leq n\) such that \(\beta = \alpha_k, \gamma = \alpha_n\), and so \(\beta, \gamma \in K_n\). Since \(K_n\) is a field, we have
\[
\beta^{-1}, \gamma^{-1}, \beta + \gamma, \beta \gamma \in K_n \subseteq \mathbb{A}.
\]
Therefore \(\mathbb{A}\) is a field. \(\Box\)
An immediate implication of Theorem 5.1.3 is that a sum, a difference, a product, or a quotient of two algebraic numbers is again an algebraic number. This not always true for transcendental numbers, which is what we show next.

**Lemma 5.1.4.** A sum or product of an algebraic number and a transcendental number is transcendental.

**Proof.** Let \( \alpha \in \mathbb{C} \) be algebraic and \( \beta \in \mathbb{C} \) transcendental. Then \( -\alpha \) and \( \alpha^{-1} \) are algebraic. Suppose that \( \alpha + \beta \) and \( \alpha \beta \) are algebraic. Since sum and product of algebraic numbers are algebraic, we must have

\[
\beta = (\alpha + \beta) + (-\alpha) = (\alpha \beta)\alpha^{-1} \in \mathbb{A},
\]

which is a contradiction. Hence \( \alpha + \beta \) and \( \alpha \beta \) must be transcendental. \( \square \)

**Remark 5.1.1.** A consequence of Lemma 5.1.4 is that, given one transcendental number \( \beta \), we can produce infinitely many (but countably many) transcendental numbers:

\[
\alpha \pm \beta, \alpha \beta, \alpha \beta^{-1} \forall 0 \neq \alpha \in \mathbb{A},
\]

Take, for instance, \( \beta \) to be a Liouville number.

**Example 5.1.2.** Let \( \alpha \in \mathbb{C} \) be algebraic and \( \beta \in \mathbb{C} \) transcendental. Then \( \alpha \beta, \alpha + \beta \) are transcendental by Lemma 5.1.4. On the other hand, \( \alpha = (\alpha + \beta) - \beta = \frac{\alpha\beta}{\beta} \) is algebraic. Hence \( \mathbb{T} \) is not a field.

The notion of algebraicity can also be generalized over extensions of \( \mathbb{Q} \) as follows. Let \( K \subset \mathbb{C} \) be a field and \( \alpha \in \mathbb{C} \). We say that \( \alpha \) is algebraic over \( K \) if there exists a polynomial \( f(x) \) with coefficients in \( K \) such that \( f(\alpha) = 0 \). Then the minimal polynomial of \( \alpha \) over \( K \), denoted \( m_{\alpha,K}(x) \) is the monic such polynomial of smallest degree. We say that a polynomial \( f(x) \in K[x] \) is irreducible over \( K \) if whenever \( f(x) \) is factored as

\[
f(x) = g(x)h(x)
\]

with \( g(x), h(x) \in K[x] \), then either \( g(x) \) or \( h(x) \) is a constant. By the same logic as in Problem 4.3, \( m_{\alpha,K}(x) \) is irreducible over \( K \) and \( m_{\alpha,K}(x) \mid p(x) \) for every \( p(x) \in K[x] \) such that \( p(\alpha) = 0 \).

To conclude this section, we introduce the notion of algebraic independence.

**Definition 5.1.3.** Let \( \alpha, \beta \in \mathbb{C} \) be transcendental numbers. Then, as we know from Corollary 5.1.2,

\[
[\mathbb{Q}(\alpha) : \mathbb{Q}] = [\mathbb{Q}(\beta) : \mathbb{Q}] = \infty.
\]

These numbers are called algebraically independent if

\[
[\mathbb{Q}(\alpha, \beta) : \mathbb{Q}(\alpha)] = [\mathbb{Q}(\alpha, \beta) : \mathbb{Q}(\beta)] = \infty.
\]

More generally, a collection of transcendental numbers \( \alpha_1, \ldots, \alpha_n \) is algebraically independent if the degree of \( \mathbb{Q}(\alpha_1, \ldots, \alpha_n) \) over \( \mathbb{Q}(S) \), where \( S \) is any proper subcollection of \( \alpha_1, \ldots, \alpha_n \), is equal to infinity. If \( K \) is a subfield of \( \mathbb{C} \), then its transcendence degree, denoted \( \operatorname{trdeg} K \), is the cardinality of a maximal (with respect to size) collection of algebraically independent elements in \( K \).
Notice that no subcollection of each infinite collection of transcendental numbers mentioned in Remark 5.1.1 is algebraically independent. In other words, while we can construct infinitely many transcendental numbers given one, it is not so easy to construct algebraically independent transcendental numbers.
5.2. Number fields and rings of integers

We now need to introduce some further language of algebraic number theory.

**Definition 5.2.1.** Let \( \alpha \in \mathbb{A} \), then the *algebraic conjugates* of \( \alpha \) (also often called just *conjugates*) are all the roots of its minimal polynomial \( m_\alpha(x) \).

A polynomial \( f(x) \in \mathbb{C}[x] \) of degree \( n \) is called *separable* if all of its \( n \) roots in \( \mathbb{C} \) are distinct.

**Lemma 5.2.1.** Let \( f(x) \in \mathbb{Z}[x] \) be an irreducible polynomial. Then it is separable.

**Proof.** Suppose

\[
f(x) = \sum_{k=0}^{n} a_k x^k \in \mathbb{Z}[x],
\]

and assume that \( \alpha \in \mathbb{C} \) is a root of \( f(x) \). By Problem 4.3, \( m_\alpha(x) | f(x) \), which means that \( f(x) = m_\alpha(x) \), since \( f(x) \) is irreducible. Then there exists some polynomial \( g(x) \in \mathbb{C}[x] \) such that

\[
f(x) = (x - \alpha)^\ell g(x),
\]

where \( \ell \geq 1 \) is the multiplicity of \( \alpha \) as a root of \( f(x) \) and \( g(\alpha) \neq 0 \). We want to prove that \( \ell = 1 \). Arguing towards a contradiction, suppose that \( \ell > 1 \). Let \( f'(x) \) be the formal derivative of \( f(x) \), i.e.

\[
f'(x) = \sum_{k=1}^{n} k a_k x^{k-1} \in \mathbb{Z}[x].
\]

The standard differentiation product rule applies, and so

\[
f'(x) = \ell(x - \alpha)^{\ell-1} g(x) + (x - \alpha)^{\ell} g'(x).
\]

Hence \( f'(\alpha) = 0 \), and so by Problem 4.3, \( f(x) \) is separable. On the other hand, degree of \( f'(x) \) is less than degree of \( f(x) \), while \( f'(x) \) is not identically 0. This is a contradiction, therefore \( \ell = 1 \). Since this is true for every root of \( f(x) \), it must be separable. \( \square \)

This lemma has an immediate corollary.

**Corollary 5.2.2.** Let \( \alpha \in \mathbb{C} \) be an algebraic number. Then all of its algebraic conjugates are distinct.

**Proof.** Since \( m_\alpha(x) \), the minimal polynomial of \( \alpha \) is irreducible (Problem 4.3), it must be separable by Lemma 5.2.1, and hence \( \alpha \) and its conjugates are all distinct. \( \square \)

**Definition 5.2.2.** A finite algebraic extension of \( \mathbb{Q} \) is called a *number field*. In other words, a number field \( K \) is a subfield of \( \mathbb{C} \) such that \( |K : \mathbb{Q}| < \infty \) and every element of \( K \) is algebraic.

It is clear from the above definition that every number field \( K \) is contained in \( \mathbb{A} \). Furthermore, if \( \alpha \in \mathbb{A} \), then \( \mathbb{Q}(\alpha) \) is an algebraic extension of \( \mathbb{Q} \), and \( |\mathbb{Q}(\alpha) : \mathbb{Q}| = \deg(\alpha) < \infty \), hence \( \mathbb{Q}(\alpha) \) is a number field. An element \( \alpha \) in a number field \( K \) is called a *primitive element* if \( K = \mathbb{Q}(\alpha) \), i.e. if \( \alpha \) generates \( K \) over \( \mathbb{Q} \). In fact, all number fields contain a primitive element.
Theorem 5.2.3 (Primitive Element Theorem). Let $K$ be a number field. Then there exists $\alpha \in K$ such that $K = \mathbb{Q}(\alpha)$.

Proof. Since $K$ is a finite algebraic extension of $\mathbb{C}$, there must exist a finite collection of algebraic numbers $\alpha_1, \ldots, \alpha_n \in K$ such that $K = \mathbb{Q}(\alpha_1, \ldots, \alpha_n)$ (Problem 5.10). Let $K_1 = \mathbb{Q}(\alpha_1)$, $K_2 = \mathbb{Q}(\alpha_1, \alpha_2) = K_1(\alpha_2)$, $\ldots$, $K_n = \mathbb{Q}(\alpha_1, \ldots, \alpha_{n-1}, \alpha_n) = K_{n-1}(\alpha_n)$. We can assume that no $K_m$ equal to $K_{m+1}$, since otherwise we do not need $\alpha_{m+1}$ in the generating set. Hence we have

$$\mathbb{Q} \subset K_1 \subset K_2 \subset \cdots \subset K_{n-1} \subset K_n = K.$$

Notice that it is sufficient for us to show that there exists $\beta_1 \in K$ such that $K_2 = \mathbb{Q}(\alpha_1, \alpha_2) = K(\beta_1)$: if this the case, then applying the same reasoning, we establish that

$$K_3 = K_2(\alpha_3) = \mathbb{Q}(\beta_1, \alpha_3) = \mathbb{Q}(\beta_2)$$

for some $\beta_2 \in K$, and continuing in the same manner confirm that $K = K_n = \mathbb{Q}(\beta_{n-1})$ for some $\beta_{n-1} \in K$.

Let $\deg(\alpha_1) = d$, $\deg(\alpha_2) = e$, and let

$$\alpha_1 = \alpha_{11}, \alpha_{12}, \ldots, \alpha_{1d} \text{ and } \alpha_2 = \alpha_{21}, \alpha_{22}, \ldots, \alpha_{2e}.$$

be algebraic conjugates of $\alpha_1$ and $\alpha_2$, respectively. Since $m_{\alpha_1}(x)$ and $m_{\alpha_2}(x)$ in $\mathbb{Z}[x]$ are irreducible, they must be separable by Lemma 5.2.1 above, and hence all $\alpha_1$’s and all $\alpha_2$’s are distinct. This means that for each $1 \leq n \leq d$, $1 < m \leq e$ the equation

$$(5.5) \quad \alpha_{1n} + t\alpha_{2m} = \alpha_{11} + t\alpha_{21}$$

has at most one solution $t$ in $\mathbb{Q}$ (a solution $t$ in $\mathbb{C}$ always exists, but it may not be in $\mathbb{Q}$). There are only finitely many equations (5.5), each having at most one solution, and hence we can choose $0 \neq c \in \mathbb{Q}$ which is not one of these solutions, then

$$\alpha_{1n} + c\alpha_{2m} \neq \alpha_{11} + c\alpha_{21}$$

for any $1 \leq n \leq d$, $1 < m \leq e$. Let

$$\beta_1 = \alpha_1 + c\alpha_2,$$

then $\beta_1 \neq \alpha_{1n} + c\alpha_{2m}$ for any $1 \leq n \leq d$, $1 < m \leq e$. We will now prove that $\mathbb{Q}(\beta_1) = \mathbb{Q}(\alpha_1, \alpha_2)$. It is clear that $\mathbb{Q}(\beta_1) \subseteq \mathbb{Q}(\alpha_1, \alpha_2)$, so we only need to show that $\mathbb{Q}(\alpha_1, \alpha_2) \subseteq \mathbb{Q}(\beta_1)$. For this, it is sufficient to prove that $\alpha_2 \in \mathbb{Q}(\beta_1)$, since then $\alpha_1 = \beta_1 - c\alpha_2 \in \mathbb{Q}(\beta_1)$, and hence $\mathbb{Q}(\alpha_1, \alpha_2) \subseteq \mathbb{Q}(\beta_1)$. Notice that

$$m_{\alpha_1}(\beta_1 - c\alpha_2) = m_{\alpha_1}(\alpha_1) = 0.$$

In other words, $\alpha_2$ is a zero of the polynomial

$$p(x) := m_{\alpha_1}(\beta_1 - cx),$$

which has coefficients in $\mathbb{Q}(\beta_1)$. On the other hand, $\alpha_2$ is also a root of its minimal polynomial $m_{\alpha_2}(x)$. The two polynomials $p(x)$ and $m_{\alpha_2}(x)$ have only one common root. Indeed, if $\xi \in \mathbb{C}$ is such that

$$p(\xi) = m_{\alpha_1}(\beta_1 - c\xi) = m_{\alpha_2}(\xi) = 0,$$

then $\xi$ must be one of $\alpha_{21}, \ldots, \alpha_{2e}$ and $\beta_1 - c\xi$ one of $\alpha_{11}, \ldots, \alpha_{1d}$, i.e., for some $1 \leq n \leq d$, $1 < m \leq e$,

$$\xi = \alpha_{2m} \text{ and } \beta_1 - c\xi = \beta_1 - c\alpha_{2m} = \alpha_{1n},$$

for some $1 \neq n \neq m$. Therefore, $\alpha_2$ is algebraic conjugate of $\alpha_1$.
which means that
\[ \beta_1 = \alpha_{1n} + c\alpha_{2m} = \alpha_{11} + c\alpha_{21}. \]
This contradicts our choice of \( c \) unless \( n = m = 1 \).

Now let \( h(x) \) be a minimal polynomial of \( \alpha_2 \) over \( \mathbb{Q}(\beta_1) \). Since \( p(x) \) and \( m_{\alpha_2}(x) \) have coefficients in \( \mathbb{Q}(\beta_1) \) and vanish at \( \alpha_2 \), they must both be divisible by \( h(x) \) over \( \mathbb{Q}(\beta_1) \). This means that every root of \( h(x) \) would be a common root of \( p(x) \) and \( m_{\alpha_2}(x) \), but we know that they have precisely one root in common. This means that \( h(x) \) can have only one root, and hence is of degree 1. Thus
\[ h(x) = x - \alpha_2, \]
which means that \( \alpha_2 \in \mathbb{Q}(\beta_1) \). This completes the proof. \( \Box \)

An algebraic number \( \alpha \) is called an algebraic integer if its minimal polynomial \( m_{\alpha}(x) \in \mathbb{Z}[x] \) is monic, i.e. its leading coefficient is equal to 1. The set of all algebraic integers in a number field \( K \) is usually denoted by \( \mathcal{O}_K \). For instance, \( \mathcal{O}_\mathbb{Q} = \mathbb{Z} \) (Problem 5.9). Let us define the set of all algebraic integers
\[ \mathcal{I} = \{ \alpha \in \mathbb{A} : m_{\alpha}(x) \text{ is monic} \}. \]

Let \( \alpha \in \mathcal{I} \) and let \( \deg(\alpha) = d \). Define
\[ \mathbb{Z}[\alpha] := \left\{ \sum_{n=0}^{d-1} a_n \alpha^n : a_0, \ldots, a_{d-1} \in \mathbb{Z} \right\}. \]

**Lemma 5.2.4.** Let \( \alpha \in \mathcal{I} \) have degree \( d \). Then \( \mathbb{Z}[\alpha] \) is a commutative ring with identity under the usual addition and multiplication operations on complex numbers, which contains \( \mathbb{Z} \). Rings like this are called ring extensions of \( \mathbb{Z} \).

**Proof.** The argument here bears some similarity with the proof of Theorem 5.1.1 above. It is clear that \( \mathbb{Z} \subseteq \mathbb{Z}[\alpha] \), and hence \( 0, 1 \in \mathbb{Z}[\alpha] \). Also, if \( \beta = \sum_{n=0}^{d-1} b_n \alpha^n \in \mathbb{Z}[\alpha] \), then \( -\beta = \sum_{n=0}^{d-1} (-b_n) \alpha^n \in \mathbb{Z}[\alpha] \). Hence we only need to prove that for every \( \beta, \gamma \in \mathbb{Z}[\alpha] \), \( \beta + \gamma, \beta \gamma \in \mathbb{Z}[\alpha] \). Let
\[ \beta = \sum_{n=0}^{d-1} b_n \alpha^n, \quad \gamma = \sum_{n=0}^{d-1} c_n \alpha^n. \]

Then
\[ \beta + \gamma = \sum_{n=0}^{d-1} (b_n + c_n) \alpha^n \in \mathbb{Z}[\alpha]. \]

Since \( \alpha \in \mathcal{I} \) of degree \( d \), its minimal polynomial is monic of degree \( d \), say
\[ m_{\alpha}(x) = \alpha^d + \sum_{n=0}^{d-1} a_n x^n, \]
and \( m_{\alpha}(\alpha) = 0 \), meaning that
\[ \alpha^d = -\sum_{n=0}^{d-1} a_n \alpha^n. \]

(5.6)
Now we have:
\[ \beta \gamma = \sum_{n=0}^{d-1} \sum_{m=0}^{d-1} b_n c_m \alpha^{n+m}, \]
and (5.6) can be used to express powers of \( \alpha \) higher than \((d - 1)\)-st as linear combinations of lower powers of \( \alpha \) with rational integer coefficients, hence ensuring that \( \beta \gamma \) is a linear combination of the terms \( 1, \alpha, \ldots, \alpha^{d-1} \) with coefficients in \( \mathbb{Z} \). This means that \( \beta \gamma \in \mathbb{Z}[\alpha] \) and completes the proof of the lemma. \( \square \)

Problem 5.11 guarantees that \( 1, \alpha, \ldots, \alpha^{d-1} \) is a maximal linearly independent collection of powers of \( \alpha \) over \( d \), and we know that it spans \( \mathbb{Z}[\alpha] \). Hence this is a basis for \( \mathbb{Z}[\alpha] \) over \( \mathbb{Z} \), and thus \( \mathbb{Z}[\alpha] \) is a lattice of rank \( d \).

**Lemma 5.2.5.** Let \( \alpha \in \mathbb{C} \) be such that the additive abelian group generated by all powers of \( \alpha \) is in fact finitely generated. Then \( \alpha \in \mathbb{I} \).

**Proof.** Let \( G \) be the additive abelian group generated by all powers of \( \alpha \), i.e.

\[
G = \left\{ \sum_{n=0}^{k} a_n \alpha^n : k \in \mathbb{N}_0, \; a_0, \ldots, a_k \in \mathbb{Z} \right\}.
\]

Assume that \( G \) is finitely generated and let \( v_1, \ldots, v_m \) be a generating set for \( G \). Since each \( v_n \) is a polynomial in \( \alpha \), there exists a positive integer \( \ell \) which is the maximal power of \( \alpha \) present in the representations of \( v_1, \ldots, v_m \). Then \( G \) is generated by \( 1, \alpha, \ldots, \alpha^\ell \). Since \( \alpha^{\ell+1} \in G \), there must exist \( a_0, \ldots, a_\ell \in \mathbb{Z} \) such that

\[
\alpha^{\ell+1} = \sum_{n=0}^{\ell} a_n \alpha^n,
\]

which means that \( \alpha \) is a root of the polynomial

\[
p(x) = x^{\ell+1} - \sum_{n=0}^{\ell} a_n x^n \in \mathbb{Z}[x].
\]

Therefore we must have \( m_\alpha(x) \mid p(x) \). Since \( p(x) \) is a monic polynomial, it must be true that \( m_\alpha(x) \) is also monic. Hence \( \alpha \in \mathbb{I} \). \( \square \)

**Theorem 5.2.6.** \( \mathbb{I} \) is a commutative ring with identity under the usual addition and multiplication of complex numbers.

**Proof.** We only need to prove that for any \( \alpha, \beta \in \mathbb{I}, \alpha + \beta \) and \( \alpha \beta \) are in \( \mathbb{I} \). Notice that \( \alpha + \beta \) and \( \alpha \beta \) can be expressed as integral linear combinations of elements of the form \( \alpha^m \beta^n \) for some nonnegative integers \( m, n \), which means that

\[
\alpha + \beta, \alpha \beta \in G := \text{span}_{\mathbb{Z}} \{ \alpha^m \beta^n : m, n \in \mathbb{Z}_{\geq 0} \} \subset \mathbb{C}.
\]

This \( G \) is a subgroup of \( \mathbb{C} \) under the usual addition of complex numbers, and hence is an additive abelian group (Problem 5.12). Since \( \alpha \) and \( \beta \) are algebraic integers, we know that \( \mathbb{Z}[\alpha] \) and \( \mathbb{Z}[\beta] \) are generated by only finitely many powers of \( \alpha \) and \( \beta \), respectively, say, it is \( 1, \alpha, \ldots, \alpha^k \) and \( 1, \beta, \ldots, \beta^\ell \). Then \( G \) is generated by all expressions of the form \( \alpha^m \beta^n, 0 \leq m \leq k, 0 \leq n \leq \ell \) as an additive abelian group. Therefore \( G \) must also be finitely generated. We now need to use a standard property of finitely generated abelian groups, the proof of which we postpone to Appendix A.
Fact 5.2.1. Let $G$ be a finitely generated additive abelian group, i.e., there exist $v_1, \ldots, v_k \in G$ such that for every $x \in G$,

$$x = \sum_{n=1}^{k} a_n v_n$$

for some $a_1, \ldots, a_k \in \mathbb{Z}$. Let $H$ be a subgroup of $G$. Then $H$ is also finitely generated.

Since additive groups generated by all powers of $\alpha + \beta$ and $\alpha \beta$, respectively, are subgroups of $G$, they must also be finitely generated. Now Lemma 5.2.5 guarantees that $\alpha + \beta$ and $\alpha \beta$ must be in $\mathbb{I}$.$\square$

Notice that we can now describe the set of all algebraic integers in a number field $K$ as $\mathcal{O}_K = K \cap \mathbb{I}$. This implies that $\mathcal{O}_K$ is a ring (Problem 5.13). We now further study some properties of the ring of algebraic integers $\mathcal{O}_K$ of a number field $K$. First we observe that every element of $K$ can be expressed as a fraction $\alpha/c$, where $\alpha$ is an algebraic integer and $c$ is a rational integer.

Lemma 5.2.7. Let $K$ be a number field and $\beta \in K$. Then there exists some $c \in \mathbb{N}$ such that $c \beta \in \mathcal{O}_K$. In fact, we can take $c$ to be the leading coefficient of $m_\beta(x)$.

Proof. Let $d = \deg(\beta)$ and let

$$m_\beta(x) = \sum_{n=0}^{d} a_n x^n \in \mathbb{Z}[x]$$

with $a_d > 0$. Notice that

$$p(x) := a_d^{d-1} m_\beta(x) = \sum_{n=0}^{d} a_n a_d^{d-1-1} x^n = \sum_{n=0}^{d} a_n a_d^{d-n-1} (a_d x)^n$$

has $\beta$ as its root. Now

$$f(x) = \sum_{n=0}^{d} a_n a_d^{d-n-1} x^n = x^d + \sum_{n=0}^{d-1} a_n a_d^{d-n-1} x^n \in \mathbb{Z}[x]$$

is a monic polynomial, and $f(a_d \beta) = p(\beta) = 0$. This means that $a_d \beta \in \mathcal{O}_K$. Taking $c = a_d$ completes the proof of the lemma.$\square$

This lemma has some important corollaries.

Corollary 5.2.8. A number field $K$ can be described as

$$K = \left\{ \frac{\alpha}{\beta} : \alpha, \beta \in \mathcal{O}_K, \beta \neq 0 \right\}.$$

Hence we can refer to $K$ as the field of fractions or quotient field of $\mathcal{O}_K$.

Proof. Let

$$E := \left\{ \frac{\alpha}{\beta} : \alpha, \beta \in \mathcal{O}_K, \beta \neq 0 \right\}.$$

We need to prove that $E = K$. Lemma 5.2.7 implies that every $\beta \in K$ can be written as $\beta = \frac{a}{c}$ for some $\alpha \in \mathcal{O}_K$ and $c \in \mathbb{Z}$. Since $\mathbb{Z} \subseteq \mathcal{O}_K$, we see that $\beta \in E$,
hence $K \subseteq E$. Now suppose $\alpha/\beta = \alpha \beta^{-1} = E$. Since $\alpha, \beta \in O_K \subseteq K$, we must have $\beta^{-1} \in K$ and hence $\alpha \beta^{-1} \in K$, since $K$ is a field. Therefore $E \subseteq K$, and thus $E = K$. \qed

Theorem 5.2.3 guarantees that a number field always has a primitive element. In fact, it always has a primitive element, which is an algebraic integer.

Corollary 5.2.9. Let $K$ be a number field. Then there exists $\alpha \in O_K$ such that $K = \mathbb{Q}(\alpha)$.

Proof. Let $\beta \in K$ be a primitive element. By Lemma 5.2.7, there exists an element $c \in \mathbb{Z}$ such that $\alpha := c \beta \in O_K$. Since clearly $\mathbb{Q}(c\beta) = \mathbb{Q}(\beta)$, we are done. \qed

We can now define embeddings of a number field $K$ into $\mathbb{C}$. Let $K = \mathbb{Q}(\alpha)$, then

$$d := \deg(\alpha) = [K : \mathbb{Q}].$$

Recall that

$$K = \mathbb{Q}(\alpha) = \mathbb{Q}[\alpha] = \text{span}_{\mathbb{Q}}\{1, \alpha, \ldots, \alpha^{d-1}\},$$

and $1, \alpha, \ldots, \alpha^{d-1}$ are linearly independent over $\mathbb{Q}$. Let

$$\alpha = \alpha_1, \alpha_2, \ldots, \alpha_d$$

be the algebraic conjugates of $\alpha$. For each $1 \leq n \leq d$, define a map $\sigma_n : K \to \mathbb{C}$, given by

$$\sigma_n \left( \sum_{m=0}^{d-1} a_m \alpha^m \right) = \sum_{m=0}^{d-1} a_m \alpha_n^m,$$

for each $\sum_{m=0}^{d-1} a_m \alpha^m \in K$. From Problem 5.14 we know that each such $\sigma_n$ is an injective field homomorphism, so $K \cong \sigma_n(K)$ for each $1 \leq n \leq d$, and

$$Q = \{ \beta \in K : \sigma_n(\beta) = \beta \forall 1 \leq n \leq d \}.$$  

These embeddings $\sigma_1, \ldots, \sigma_d$ are, in fact, the only possible embeddings of $K$ into $\mathbb{C}$.

Lemma 5.2.10. Let $K = \mathbb{Q}(\alpha)$ be a number field of degree $d$ over $\mathbb{Q}$. Let $\tau : K \to \mathbb{C}$ be an embedding, i.e. an injective field homomorphism. Then $\tau$ is one of the embeddings $\sigma_1, \ldots, \sigma_d$ as defined in (5.7).

Proof. First we will prove that $\tau(c) = c$ for each $c \in \mathbb{Q}$. Since $\tau$ is a field homomorphism, we must have $\tau(1) = 1$, and for each $a/b \in \mathbb{Q}$,

$$\tau(a/b) = \tau(a)\tau(b)^{-1} = a\tau(1)(b\tau(1))^{-1} = a/b.$$

Since $[K : \mathbb{Q}] = d$, we know that $\deg(\alpha) = d$, and so

$$K = \mathbb{Q}[\alpha] = \text{span}_\mathbb{Q}\{1, \alpha, \ldots, \alpha^{d-1}\}.$$  

Let $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_d$ be the algebraic conjugates of $\alpha$. Let $\beta = \sum_{n=0}^{d-1} c_n \alpha^n \in K$. Since $\tau$ is a field homomorphism,

$$\tau(\beta) = \sum_{n=0}^{d-1} \tau(c_n)\tau(\alpha)^n = \sum_{n=0}^{d-1} c_n \tau(\alpha)^n.$$
Hence we only need to show that \( \tau(\alpha) = \alpha_n \) for some \( 1 \leq n \leq d \). Let

\[
m_{\alpha}(x) = \sum_{m=0}^{d} b_m x^m \in \mathbb{Z}[x],
\]

be the minimal polynomial of \( \alpha \). Then

\[
m_{\alpha}(\alpha) = \sum_{m=0}^{d} b_m \alpha^m = 0,
\]

and so

\[
0 = \sum_{m=0}^{d} b_m \tau(\alpha)^m = m_{\alpha}(\tau(\alpha)).
\]

Hence \( \tau(\alpha) \) is a root of \( m_{\alpha}(x) \), which means that \( \tau(\alpha) = \alpha_n \) for some \( 1 \leq n \leq d \). Therefore \( \tau = \sigma_n \) for some \( \sigma_n \) as in (5.7). This completes the proof.

If \( K = \sigma_n(K) \) for each \( 1 \leq n \leq d \), then the number field \( K \) is called Galois. In this case, the set

\[
G := \{ \sigma_1, \ldots, \sigma_d \}
\]

is a group under the operation of function composition (Problem 5.15). It is called the Galois group of \( K \) over \( \mathbb{Q} \), where \( \mathbb{Q} \) is precisely the fixed field of \( G \), as follows from Problem 5.14. In this case, elements of \( G \) are called automorphisms of \( K \) over \( \mathbb{Q} \).

**Definition 5.2.3.** Given a \( \mathbb{Q} \)-basis \( \alpha_1, \ldots, \alpha_d \in K \), its discriminant is defined as

\[
\Delta(\alpha_1, \ldots, \alpha_d) := (\det(\sigma_n(\alpha_k))_{1 \leq n, k \leq d})^2,
\]

where \( d = [K : \mathbb{Q}] \).

We will now prove an important property of the discriminant.

**Lemma 5.2.11.** Let \( \alpha_1, \ldots, \alpha_d \in K \) be a \( \mathbb{Q} \)-basis. Then the discriminant

\[
\Delta(\alpha_1, \ldots, \alpha_d) \in \mathbb{Q}.
\]

Further, if \( \alpha_1, \ldots, \alpha_d \in \mathcal{O}_K \), then \( \Delta(\alpha_1, \ldots, \alpha_d) \in \mathbb{Z} \).

**Proof.** Let \( \theta \in K \) be such that \( K = \mathbb{Q}(\theta) \), then degree of \( \theta \) is equal to \( d \) and \( 1, \theta, \ldots, \theta^{d-1} \) is a \( \mathbb{Q} \)-basis for \( K \), called a power basis. Hence there must exist rational numbers \( c_{11}, \ldots, c_{dd} \) such that

\[
\begin{pmatrix}
\alpha_1 \\
\vdots \\
\alpha_d
\end{pmatrix} =
\begin{pmatrix}
c_{11} & \cdots & c_{1d} \\
\vdots & \ddots & \vdots \\
c_{d1} & \cdots & c_{dd}
\end{pmatrix}
\begin{pmatrix}
1 \\
\vdots \\
\theta^{d-1}
\end{pmatrix},
\]

i.e. \( C = (c_{mn})_{1 \leq m, n \leq d} \) is a rational change of basis matrix. Then by Problem 5.19

\[
\Delta(\alpha_1, \ldots, \alpha_d) = \det(C)^2 \Delta(1, \ldots, \theta^{d-1}),
\]

and thus it is sufficient to prove that \( \Delta(1, \ldots, \theta^{d-1}) \in \mathbb{Q} \). Let \( \sigma_1, \ldots, \sigma_d \) be the embeddings of \( K \) into \( \mathbb{C} \) and \( \theta_1, \ldots, \theta_d \) the conjugates of \( \theta \), i.e. \( \theta_k = \sigma_k(\theta) \). Then

\[
\Delta := \Delta(1, \ldots, \theta^{d-1}) = \left\{ \det \begin{pmatrix}
1 & \theta_1 & \cdots & \theta_1^{d-1} \\
1 & \theta_2 & \cdots & \theta_2^{d-1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & \theta_d & \cdots & \theta_d^{d-1}
\end{pmatrix} \right\}^2.
\]
Notice that each $\sigma_k$ permutes the conjugates $\theta_1, \ldots, \theta_d$, which means that the action of each $\sigma_k$ only permutes the rows of the matrix above leaving the square of the determinant unchanged. Thus $\sigma_k(\Delta) = \Delta$ for every $1 \leq k \leq d$. Now, Problem 5.20 implies that any element of $K$ that is fixed by all embeddings must in fact be in $\mathbb{Q}$, thus $\Delta \in \mathbb{Q}$, and so $\Delta(\alpha_1, \ldots, \alpha_d) \in \mathbb{Q}$ for any $\mathbb{Q}$-basis $\alpha_1, \ldots, \alpha_d$ of $K$.

Assume additionally that $\alpha_1, \ldots, \alpha_d \in O_K$, i.e. they are all algebraic integers. From definition of the discriminant it is clear that $\Delta(\alpha_1, \ldots, \alpha_d)$ in this case must also be an algebraic integer. Hence $\Delta(\alpha_1, \ldots, \alpha_d)$ is in $\mathbb{Q}$ and in $I$, but $\mathbb{Q} \cap I = \mathbb{Z}$. Thus $\Delta(\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}$.

We now use the discriminant to prove that the ring of integers $O_K$ in a number field $K$ of degree $d$ over $\mathbb{Q}$ is a lattice of rank $d$, i.e. its elements can be expressed as integer linear combinations of a collection of $d$ $\mathbb{Q}$-linearly independent elements. A ring with this property is called an order in $K$, and hence we are about to show that $O_K$ is an order in $K$. Notice that if $K = \mathbb{Q}(\theta)$ for some algebraic $\theta$, then $\mathbb{Z}[\theta] \subseteq O_K$ is also an order in $K$, however $O_K$ is a maximal order with respect to inclusion (we will not prove it here), and $\mathbb{Z}[\theta]$ is a maximal order precisely when $O_K = \mathbb{Z}[\theta]$.

**Theorem 5.2.12.** Let $K$ be a number field of degree $d$ over $\mathbb{Q}$. Then the ring $O_K$ is a lattice of rank $d$, i.e. there exists a collection of $\mathbb{Q}$-linearly independent elements $\alpha_1, \ldots, \alpha_d \in O_K$ such that

$$O_K = \left\{ \sum_{n=1}^{d} a_n \alpha_n : a_1, \ldots, a_d \in \mathbb{Z} \right\}.$$

**Proof.** Let $\alpha_1, \ldots, \alpha_d \in K$ be a $\mathbb{Q}$-basis for $K$. By Lemma 5.2.7 we know that there exist $c_1, \ldots, c_d$ such that $c_1 \alpha_1, \ldots, c_d \alpha_d \in O_K$. Thus the set of linearly independent collections of $d$ elements of $O_K$ is not empty. The discriminant of any such collection is an integer, hence let us choose such a collection $\beta_1, \ldots, \beta_d$ with the smallest $|\Delta(\beta_1, \ldots, \beta_d)|$.

We will now prove that

$$O_K = \left\{ \sum_{n=1}^{d} a_n \beta_n : a_1, \ldots, a_d \in \mathbb{Z} \right\}.$$

Suppose this is not true, then there exists some $x \in O_K$, which is not an integer linear combination of $\beta_1, \ldots, \beta_d$. Since $\beta_1, \ldots, \beta_d$ for a $\mathbb{Q}$-basis for $K$, it still must be true that

$$x = a_1 \beta_1 + \cdots + a_d \beta_d$$

for some $a_1, \ldots, a_d \in \mathbb{Q}$, which are not all in $\mathbb{Z}$. Assume, for instance, $a_1$ is not an integer. Let $q \in (0,1)$ be such that $a_1 - q \in \mathbb{Z}$. Then $(a_1 - q) \beta_1 \in O_K$, and since $x \in O_K$, we have

$$y := x - (a_1 - q) \beta_1 = q \beta_1 + \sum_{k=2}^{d} a_k \beta_k \in O_K.$$
The collection of elements \( y, \beta_2, \ldots, \beta_d \in \mathcal{O}_K \) is again linearly independent, hence forms a \( \mathbb{Q} \)-basis for \( K \). Then

\[
A = \begin{pmatrix}
q & a_2 & \ldots & a_d \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1
\end{pmatrix}
\]

is the change of basis matrix from the basis \( \beta_1, \ldots, \beta_d \) to the basis \( y, \beta_2, \ldots, \beta_d \).

Hence, by Problem 5.19

\[
\Delta(y, \beta_2, \ldots, \beta_d) = \det(A)^2 \Delta(\beta_1, \beta_2, \ldots, \beta_d)
\]

\[
= q^2 \Delta(\beta_1, \beta_2, \ldots, \beta_d) < \Delta(\beta_1, \beta_2, \ldots, \beta_d).
\]

This contradicts the choice of \( \beta_1, \ldots, \beta_d \), and hence completes our proof.  \( \Box \)

A \( \mathbb{Z} \)-basis (i.e. a basis over \( \mathbb{Z} \)) \( \alpha_1, \ldots, \alpha_d \) for the ring of integers \( \mathcal{O}_K \) is called an **integral basis** for the number field \( K \). If \( \alpha_1, \ldots, \alpha_d \) and \( \beta_1, \ldots, \beta_d \) are two such bases, then there must exist a change of basis matrix \( A \in \text{GL}_d(\mathbb{Z}) \) between them (Problem 5.21). Therefore we have, by Problem 5.19,

\[
\Delta(\alpha_1, \ldots, \alpha_d) = \Delta(\beta_1, \ldots, \beta_d).
\]

This common value of the discriminant of all the integral bases of a number field \( K \) is called the **discriminant of \( K \)**, denoted \( \Delta_K \). As we will see later, it has some special geometric significance.
5.3. Noetherian rings and factorization

We are now starting to study some properties of rings of algebraic integers like $\mathcal{O}_K$ for a number field $K$ in more details. We start with an important general theorem.

**Theorem 5.3.1.** Let $R$ be a commutative ring with identity. The following properties in $R$ are equivalent:

1. Every ideal in $R$ is finitely generated.
2. Every ascending chain of ideals
   \[ I_1 \subseteq I_2 \subseteq \cdots \]
   in $R$ stabilizes, i.e. there exists an $n$ such that $I_k = I_{k+1}$ for all $k \geq n$.
3. Given any nonempty collection of ideals $S$ in $R$, there exists an ideal $I \in S$ such that $I \nsubseteq J$ for any $J \in S$, $J \neq I$: such an $I$ is called a maximal element in $S$.

**Proof.** Suppose every ideal in $R$ is finitely generated, and let
\[ I_1 \subseteq I_2 \subseteq \cdots \]
be an ascending chain of ideals in $R$. Let $I = \bigcup_{k=1}^{\infty} I_k$, then $I$ is also an ideal in $R$ (Problem 5.22). Hence $I$ must be finitely generated, say $x_1, \ldots, x_n$ is a set of generators for $I$. Then each generator $x_k$ lies in some ideal $I_{k_0}$, and so $x_1, \ldots, x_n \in I_m$, where $m = \max_{1 \leq k \leq n} \ell_k$. Hence
\[ I = \langle x_1, \ldots, x_n \rangle \subseteq I_m \subseteq I, \]
so $I = I_m$. This means that for each $k \geq m$, $I_k = I_{k+1}$.

Now assume that every ascending chain of ideals in $R$ stabilizes. Let $S$ be a nonempty collection of ideals in $R$, and suppose that $S$ does not have a maximal element. Then for every $I \in S$ there exists some $J \in S$ such that $I \nsubseteq J$. Construct an ascending chain of ideals from $S$
\[ I_1 \subseteq I_2 \subseteq \cdots \]
by picking each $I_n$ such that $I_{n-1} \subseteq I_n$. This chain will never stabilize, which is a contradiction. Hence $S$ must have a maximal element.

Finally, suppose that every nonempty collection of ideals has a maximal element. Let $I$ be an ideal in $R$, and let $S$ be the collection of all finitely generated ideals contained in $I$. Since $\{0\} \in S$, it is not empty. Let $J$ be a maximal element in $S$, then $J \nsubseteq I$ and $J$ is finitely generated. Suppose $I$ is not finitely generated, then $J \nsubseteq I$, i.e. there exists $x \in I \setminus J$. Let $J' = \langle J, x \rangle$, then $J \nsubseteq J'$ and $J'$ is still finitely generated, so $J' \in S$. This contradicts maximality of $J$ in $S$, hence $I$ must be finitely generated. □

A commutative ring with identity satisfying the equivalent conditions of Theorem 5.3.1 is called *noetherian*. One important property of noetherian integral domains is that they allow factorization of elements into irreducibles.

**Definition 5.3.1.** Let $R$ be an integral domain. An element $u \in R$ is called a *unit* if there exists an element $v \in R$ such that $uv = 1$. The set of all units in $R$, denoted by $R^\times$, forms an abelian group under multiplication (Problem 5.16). An element $x \in R$ is called *irreducible* if whenever $x = yz$ for some $y, z \in R$ then either
y or z is a unit. Notice, in particular, that if x is irreducible, then so is ux for any unit $u \in R$.

**Theorem 5.3.2.** If $R$ is a noetherian integral domain and $x \in R$, then there exist irreducible elements $\alpha_1, \ldots, \alpha_n \in R$ such that $x = \alpha_1 \cdots \alpha_n$.

**Proof.** Suppose $u \in R$ is a unit, then $x = u(u^{-1}x)$. Notice that $\pm 1 \in R$, and hence the group of units $R^\times \neq \emptyset$. Therefore it is always possible to write $x = yz$: if in every such factorization either $y$ or $z$ is a unit, then $x$ is irreducible and we are done. If this is not the case, then there exists a factorization $x = x_1x_2$, where $x_1, x_2$ are both non-units. Now repeat this process for $x_1, x_2$, and keep repeating until the process terminates. Hence we need to show that this process does in fact terminate. Suppose not, then there exists an infinite sequence of distinct elements in $R$, call them $y_1, y_2, \ldots$ such that

$$
\cdots | y_n | y_{n-1} | \cdots | y_2 | y_1 | x,
$$

which means that there is an infinite ascending chain of ideals

$$
\langle x \rangle \subsetneq \langle y_1 \rangle \subsetneq \langle y_2 \rangle \subsetneq \cdots,
$$

which does not stabilize. This contradicts the assumption that $R$ is noetherian. Hence the process must terminate, meaning that we obtain a factorization of $x$ into irreducibles. □

We are now ready to discuss factorization of elements into irreducibles in $\mathcal{O}_K$.

**Lemma 5.3.3.** Let $K$ be a number field and $\mathcal{O}_K$ its ring of integers. Then $\mathcal{O}_K$ is noetherian.

**Proof.** We prove this lemma by showing that every ideal in $\mathcal{O}_K$ is finitely generated. By Theorem 5.2.12 we know that $\mathcal{O}_K$ is a lattice, i.e. a free abelian group. Let $I \subseteq \mathcal{O}_K$ be an ideal, then it is a subgroup of $\mathcal{O}_K$, which must therefore also be free abelian as discussed in Appendix A. Let $x_1, \ldots, x_m$ be a basis for $I$, then

$$
I = \left\{ \sum_{k=1}^{m} a_k x_k : a_1, \ldots, a_m \in \mathbb{Z} \right\} \subseteq \left\{ \sum_{k=1}^{m} a_k x_k : a_1, \ldots, a_m \in \mathcal{O}_K \right\} \subseteq I,
$$

hence $I$ is generated by $x_1, \ldots, x_m$. Thus it is finitely generated. □

**Corollary 5.3.4.** Let $K$ be a number field and $\mathcal{O}_K$ its ring of integers. Then every element in $\mathcal{O}_K$ can be factored into a product of irreducibles.

**Proof.** This is immediate by combining Theorem 5.3.2 with Lemma 5.3.3. □

Our next goal is to investigate uniqueness of factorization into irreducibles in rings of algebraic integers of number fields.

**Definition 5.3.2.** An element $x$ in an integral domain $R$ is called a prime if whenever $x \mid yz$ for some $y, z \in R$, then $x \mid y$ or $x \mid z$.

We are used to the situation of the ring of rational integers $\mathbb{Z}$, in which the group of units is $\{\pm 1\}$ and primes and irreducibles are the same (Problem 5.17). For more general rings of integers, the situation is more complicated, starting with the fact that the group of units can be larger. For instance, notice that

$$
\sqrt{2} - 1, \sqrt{2} + 1 \in \mathbb{Q}(\sqrt{2}),
$$

where $\mathbb{Q}(\sqrt{2})$ is the field of rational numbers extended by $\sqrt{2}$.
and \((\sqrt{2} - 1)(\sqrt{2} + 1) = 1\), hence they are both units. The relationship between primes and irreducibles is also not so simple.

**Lemma 5.3.5.** Let \(x \in \mathcal{O}_K\) be a prime. Then it is irreducible.

**Proof.** Suppose \(x = yz\) for some two \(y, z \in \mathcal{O}_K\). Since \(x\) is a prime, it must be true that \(x \mid y\) or \(x \mid z\), say \(x \mid y\). Then \(y = xt\) for some \(t \in \mathcal{O}_K\). Hence we have:

\[ x = xtz, \]

and multiplying this equation by \(x^{-1}\) in \(K\), we conclude that \(tz = 1\), i.e. \(z\) is a unit. Thus \(x\) is irreducible. \(\square\)

The converse of this lemma however is not always true. For example, in \(\mathcal{O}_{\mathbb{Q}(\sqrt{-5})}\) the elements 2, 3, \(1 \pm \sqrt{-5}\) are all irreducible, 2 and 3 do not divide \(1 + \sqrt{-5}\) or \(1 - \sqrt{-5}\) (Problem 5.18), however

\[(5.8) \quad 6 = 2 \times 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}),\]

and hence 2 and 3 are not primes. The underlying reason for this is the non-uniqueness of factorization into irreducibles demonstrated in (5.8). Recall that an integral domain \(R\) is called a unique factorization domain (UFD) if for any \(x \in R\) there exists a unique (up to permutation of terms and multiplication by a unit) factorization

\[ x = p_1 \cdots p_k, \]

where \(p_1, \ldots, p_k\) are irreducible elements (notice that if \(p\) is an irreducible and \(u\) is a unit, then \(up\) is also an irreducible). We are quite used to taking this property for granted, as the Fundamental Theorem of Arithmetic is nothing else but the statement that \(\mathbb{Z}\) is a UFD, however, as (5.8) demonstrates, \(\mathcal{O}_{\mathbb{Q}(\sqrt{-5})}\) is not a UFD.

**Theorem 5.3.6.** The ring \(\mathcal{O}_K\) is a UFD if and only if every irreducible in \(\mathcal{O}_K\) is a prime.

**Proof.** First suppose \(\mathcal{O}_K\) is a UFD. Let \(p \in \mathcal{O}_K\) be irreducible, and suppose that \(p \mid ab\) for some \(a, b \in \mathcal{O}_K\). Then there exists some \(c \in \mathcal{O}_K\) such that \(pc = ab\). Let

\[ a = q_1 \cdots q_k, \quad b = t_1 \cdots t_m, \quad c = s_1 \cdots s_n \]

be unique factorizations of \(a\), \(b\) and \(c\) into irreducibles, then the factorization

\[ ps_1 \cdots s_n = q_1 \cdots q_k t_1 \cdots t_m \]

is also unique. Therefore \(p\) must be one of the irreducibles \(q_1, \ldots, q_k, t_1, \ldots, t_m\), which means that \(p \mid a\) or \(p \mid b\). Hence \(p\) is prime.

In the opposite direction, assume every irreducible in \(\mathcal{O}_K\) is prime. Suppose some element \(\alpha \in \mathcal{O}_K\) has two factorizations into irreducibles, say

\[ x = p_1 \cdots p_n = q_1 \cdots q_m. \]

We want to prove that \(n = m\) and \(p_i\)'s and \(q_j\)'s are the same up to permutation. Assume \(n \geq m\), and let \(m\) be the length of the shortest factorization of \(x\) into irreducibles. We argue by induction on \(m\). If \(m = 0\), then \(x\) is a unit, and so cannot be divisible by any irreducibles, meaning that factorization is unique. Suppose now that factorization into irreducibles is unique for every element with the length of shortest factorization \(\leq m - 1\). Let us prove it for \(m\). Since \(q_m\) is a prime and

\[ q_m \mid p_1 \cdots p_n, \]
it must be true that \( q_m \) divides some \( p_j \), say \( p_n \). Thus \( p_n = uq_m \) for some \( u \in \mathcal{O}_K \), which must be a unit since \( p_n \) is irreducible. Therefore

\[
x = p_1 \cdots p_{n-1}(uq_m) = q_1 \cdots q_m.
\]

Since this equation can be viewed in the field \( K \), we have:

\[
up_1 \cdots p_{n-1} = q_1 \cdots q_{m-1},
\]

which is an element in \( \mathcal{O}_K \) with length of shortest factorization \( \leq m - 1 \). By induction hypothesis, it has unique factorization into irreducibles (up to permutation and multiplication by units), and hence so does \( x \). \qed
5.4. Norm, trace, discriminant

We will now introduce two very important functions on number fields, that will serve as essential tools in our study of properties of the rings of algebraic integers. Let $K$ be a number field of degree $d$ with embeddings $\sigma_1, \ldots, \sigma_d : K \to \mathbb{C}$. Let $\alpha \in K$, then norm of $\alpha$ in $K$ is defined as

$$N_K(\alpha) = \prod_{k=1}^{d} \sigma_k(\alpha),$$

and trace of $\alpha$ in $K$ is

$$\text{Tr}_K(\alpha) = \sum_{k=1}^{d} \sigma_k(\alpha).$$

It follows directly from these definitions that norm is a multiplicative function and trace is linear over $\mathbb{Q}$, i.e. for any $\alpha, \beta \in K$,

$$N_K(\alpha\beta) = c \ N_K(\alpha) N_K(\beta), \quad \text{Tr}_K(a\alpha + b\beta) = a \text{Tr}_K(\alpha) + b \text{Tr}_K(\beta)$$

for any $a, b, c \in \mathbb{Q}$ (Problem 5.23). There are some additional important properties of norm and trace that we will establish here.

**Lemma 5.4.1.** Let $\alpha \in K$, then $N_K(\alpha), \text{Tr}_K(\alpha) \in \mathbb{Q}$. Further, if $\alpha \in \mathcal{O}_K$, then $N_K(\alpha), \text{Tr}_K(\alpha) \in \mathbb{Z}$.

**Sketch of Proof.** Let $L = \mathbb{Q}(\alpha)$, then $L$ is a subfield of $K$ of degree $n$ over $\mathbb{Q}$, so that, by Problem 5.3,

$$d := [K : \mathbb{Q}] = [K : L][L : \mathbb{Q}] = (d/n)n,$$

where $[K : L] = d/n$. Let $\tau_1, \ldots, \tau_n$ be embeddings of $L$ and $\sigma_1, \ldots, \sigma_d$ embeddings of $K$. Then $\sigma_i$'s restricted to $L$ must be equal to $\tau_j$'s, and $\tau_j$'s extend to $\sigma_i$'s on $K$. In fact, every $\tau_j$ extends to the same number of $\sigma_i$'s, namely to $d/n$ of them, and no two different $\tau_j$'s can extend to the same $\sigma_i$. Therefore

$$N_K(\alpha) = \prod_{i=1}^{d} \sigma_i(\alpha) = \prod_{j=1}^{n} \tau_j(\alpha)^{d/n} = \left( \prod_{j=1}^{n} \tau_j(\alpha) \right)^{d/n},$$

$$\text{Tr}_K(\alpha) = \sum_{i=1}^{d} \sigma_i(\alpha) = \sum_{j=1}^{n} \left( \frac{d}{n} \tau_j(\alpha) \right) = \frac{d}{n} \sum_{j=1}^{n} \tau_j(\alpha).$$

Let $m_\alpha(x) \in \mathbb{Z}[x]$ be the minimal polynomial of $\alpha$ over $\mathbb{Z}$, then $m_\alpha(x)$ can be factored as

$$m_\alpha(x) = \sum_{j=0}^{n} c_jx^j = c_n(x - \tau_1(\alpha)) \cdots (x - \tau_n(\alpha)),$$

where $c_0, \ldots, c_n \in \mathbb{Z}$ and $c_n = 1$ if and only if $\alpha \in \mathcal{O}_K$. Hence we see that

$$c_n \prod_{j=1}^{n} \tau_j(\alpha) = c_0, \quad c_n \sum_{j=1}^{n} \tau_j(\alpha) = -c_{n-1}.$$

The result follows. □

**Corollary 5.4.2.** An element $\alpha \in \mathcal{O}_K$ is a unit if and only if $N_K(\alpha) = \pm 1$. On the other hand, if $N(\alpha) = p$, a rational prime, then $\alpha$ is irreducible.
Proof. Suppose \( \alpha \in \mathcal{O}_K \) is a unit. Then there exists \( \alpha^{-1} \in \mathcal{O}_K \) such that \( \alpha \alpha^{-1} = 1 \). Taking norms of both sides of this equation and using the fact that norm is multiplicative, we have:

\[
\mathcal{N}_K(\alpha \alpha^{-1}) = \mathcal{N}_K(\alpha) \mathcal{N}_K(\alpha^{-1}) = 1.
\]

By Lemma 5.4.1, \( \mathcal{N}_K(\alpha), \mathcal{N}_K(\alpha^{-1}) \in \mathbb{Z} \), hence we must have

\[
\mathcal{N}_K(\alpha) = \mathcal{N}_K(\alpha^{-1}) = \pm 1.
\]

On the other hand, suppose \( \mathcal{N}_K(\alpha) = \pm 1 \). There certainly exists \( \alpha^{-1} \in \mathcal{O}_K \): it is our goal to prove that \( \alpha^{-1} \in \mathcal{O}_K \). We have that

\[
1 = \mathcal{N}_K(\alpha) = \mathcal{N}_K(\alpha) \mathcal{N}_K(\alpha^{-1}) = \mathcal{N}_K(\alpha^{-1}).
\]

Since \( \alpha \in \mathcal{O}_K \) is of norm \( \pm 1 \), its minimal polynomial is of the form

\[
p(x) = x^n + \sum_{k=1}^{n-1} c_k x^k \pm 1
\]

for some \( c_1, \ldots, c_{n-1} \in \mathbb{Z} \). Then

\[
0 = p(\alpha) = \alpha^{-n} \left( \alpha^n + \sum_{k=1}^{n-1} c_k \alpha^k \pm 1 \right) = 1 + \sum_{k=1}^{n-1} c_k \alpha^{-(n-k)} \pm \alpha^{-n},
\]

that is \( \alpha^{-1} \) is a root of the monic polynomial

\[
\pm 1 \pm \sum_{k=1}^{n-1} c_k x^{n-k} + x^n \in \mathbb{Z}[x].
\]

Therefore \( \alpha^{-1} \in \mathcal{O}_K \).

Finally, suppose \( \alpha \in \mathcal{O}_K \) is such that \( \mathcal{N}_K(\alpha) = p \), a rational prime. Suppose \( \alpha = xy \) for some \( x, y \in \mathcal{O}_K \). Then Lemma 5.4.1 implies that

\[
\mathcal{N}_K(x) \mathcal{N}_K(y) = p,
\]

meaning that one of \( x, y \) has norm \( \pm 1 \) and the other \( \pm p \), hence one of them is a unit. Since this is true for every factorization of \( \alpha \) in \( \mathcal{O}_K \), we conclude that \( \alpha \) is irreducible.

We now use the norm to prove an important property of rings \( \mathcal{O}_K \).

**Theorem 5.4.3.** Let \( P \) be a prime ideal in \( \mathcal{O}_K \). Then \( P \) is maximal.

**Proof.** Let \( \alpha \in P \), and let

\[
\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n
\]

be algebraic conjugates of \( \alpha \). Notice that

\[
\alpha_2 \cdots \alpha_n = \frac{1}{\alpha} \in K.
\]

On the other hand, \( \alpha \) is an algebraic integer, therefore so are \( \alpha_2, \ldots, \alpha_n \) as well as their product. Hence \( \alpha_2 \cdots \alpha_n \) is an algebraic integer in \( K \), i.e. it is an element of \( \mathcal{O}_K \). This means that

\[
\mathcal{N}_K(\alpha) = \mathcal{N}_K(\alpha_2 \cdots \alpha_n) \in P.
\]

Lemma 5.4.1 guarantees that \( m := \mathcal{N}_K(\alpha) \in \mathbb{Z} \), hence \( m \in \mathbb{Z} \cap P \). This implies that the ideal \( m \mathcal{O}_K \) is contained in the ideal \( P \), and so

\[
|\mathcal{O}_K/P| \leq |\mathcal{O}_K/m \mathcal{O}_K|.
\]
Since $\mathcal{O}_K$ is a finitely generated abelian group, so is $\mathcal{O}_K/m\mathcal{O}_K$. Further, for any $\beta \in \mathcal{O}_K/m\mathcal{O}_K$ it $m$-th power under addition, $m\beta$, is 0. Hence $\mathcal{O}_K/m\mathcal{O}_K$ is a finitely generated abelian group in which every element has order $\leq m$ (and dividing $m$): this must be a finite group (Problem 5.24). This means $\mathcal{O}_K/P$ is finite, and since $P$ is a prime ideal, $\mathcal{O}_K/P$ is an integral domain. A finite integral domain is a field, and hence $\mathcal{O}_K/P$ is a field, which means that $P$ is a maximal ideal.

In the proof of Theorem 5.4.3 we used the fact that the quotient ring $\mathcal{O}_K/P$ is finite for a prime ideal $P$. In fact, this is true for all ideals in $\mathcal{O}_K$: notice that our argument above proving this fact did not rely on $P$ being prime. Hence we have:

**Lemma 5.4.4.** Let $I \subseteq \mathcal{O}_K$ be an ideal. Then the quotient ring $\mathcal{O}_K/I$ is finite.

We define the **norm of ideal** $I$ to be the cardinality of this finite quotient ring, i.e.

$$N_K(I) = |\mathcal{O}_K/I|.$$ 

This norm is obviously an integer, which (as we will see) generalizes the notion of the norm of an element in a number field. It plays an important role in number theory. We will prove some properties of this new norm here.

**Lemma 5.4.5.** Let $K$ be a number field of degree $d$, so $\mathcal{O}_K$ is a lattice of rank $d$. Let $I \subseteq \mathcal{O}_K$ be a nonzero ideal. Then $I$ is a lattice of rank $d$ and

$$N_K(I) = \left| \frac{\Delta(\alpha_1, \ldots, \alpha_d)}{\Delta_K} \right|^{1/2},$$

where $\alpha_1, \ldots, \alpha_d$ is an integral basis for $I$ and $\Delta_K$ is the discriminant of $K$.

**Proof.** We know that $\mathcal{O}_K$ is a lattice of rank $d$ and the ideal $I \subseteq \mathcal{O}_K$ is its sublattice. Since $\mathcal{O}_K/I$ is finite, $I$ must have the same rank as $\mathcal{O}_K$ (Problem 1.7), and so $|\mathcal{O}_K/I|$, the norm of $I$, is just the index of $I$ as a subgroup of $\mathcal{O}_K$. Then by Theorem 1.2.6,

$$N_K(I) = \frac{\det(I)}{\det(\mathcal{O}_K)} = \frac{\det(I)}{\Delta_K},$$

since the discriminant $\Delta_K$ is precisely the determinant of an integral basis matrix for $\mathcal{O}_K$. Hence we only need to compute $\det(I)$. Let $\alpha_1, \ldots, \alpha_d$ be an integral basis for $I$ and $\omega_1, \ldots, \omega_d$ an integral basis for $\mathcal{O}_K$: notice that both of these are $\mathbb{Q}$-bases for the number field $K$. Since $I$ is a sublattice of $\mathcal{O}_K$, there must exist an integral $d \times d$ matrix $A = (a_{ij})_{1 \leq i,j \leq d}$ such that

$$\alpha_i = \sum_{j=1}^d a_{ij} \omega_j$$

for each $1 \leq i \leq d$. Then, by Problem 5.19, we have:

$$\Delta(\alpha_1, \ldots, \alpha_d) = \det(A)^2 \Delta(\omega_1, \ldots, \omega_d) = \det(A)^2 \Delta_K,$$

and thus

$$N_K(I) = |\det(A)| = \left| \frac{\Delta(\alpha_1, \ldots, \alpha_d)}{\Delta_K} \right|^{1/2}.$$ 

**Corollary 5.4.6.** If $I = \langle \alpha \rangle \subseteq \mathcal{O}_K$ is a principal ideal, then

$$N_K(I) = |N_K(\alpha)|.$$
Proof. Let $\omega_1, \ldots, \omega_d$ be an integral basis for $\mathcal{O}_K$, then $\alpha \omega_1, \ldots, \alpha \omega_d$ an integral basis for $I$, and so

$$\Delta(\alpha \omega_1, \ldots, \alpha \omega_d) = (\det(\sigma_i(\alpha) \sigma_i(\omega_j))_{1 \leq i, j \leq d})^2 = \left( \prod_{i=1}^{d} \sigma_i(\alpha) \det(\sigma_i(\omega_j))_{1 \leq i, j \leq d} \right)^2 = N_K(\alpha)^2 \Delta_K.$$

Then, by Lemma 5.4.5,

$$N_K(I) = \frac{\Delta(\alpha \omega_1, \ldots, \alpha \omega_d)}{\Delta_K} \left| \frac{1}{\Delta_K} \right|^{1/2} = |N_K(\alpha)|.$$

□

We will next look in more details at the properties of ideals in the ring $\mathcal{O}_K$. The norm of an ideal will serve as an important tool, and further properties of the norm will be established later.
5.5. Fractional ideals

As we have seen, $\mathcal{O}_K$ may not necessarily have unique factorization of elements into irreducibles. On the other hand, a certain analogue of unique factorization holds for ideals in $\mathcal{O}_K$. To establish this result, we first discuss an important generalization of the notion of an ideal in number fields.

**Definition 5.5.1.** Let $K$ be a number field with ring of integers $\mathcal{O}_K$. A *fractional ideal* in $K$ is a subset

$$B = \alpha^{-1}I = \{\alpha^{-1}x : x \in I\},$$

where $\alpha \in \mathcal{O}_K$ and $I \subseteq \mathcal{O}_K$ is an ideal. Trivially, any ideal $I \subseteq \mathcal{O}_K$ is also a fractional ideal.

Let $\mathfrak{I}_K$ be the set of all fractional ideals in $K$. We can define a commutative multiplication operation on $\mathfrak{I}_K$: for $B = \alpha^{-1}I$ and $C = \beta^{-1}J$ in $\mathfrak{I}_K$,

$$BC = \{bc : b \in B, c \in C\} = (\alpha\beta)^{-1}IJ \in \mathfrak{I}_K,$$

since $\alpha\beta \in \mathcal{O}_K$ and $IJ$ is again an ideal in $\mathcal{O}_K$. Notice that for any $B = \alpha^{-1}I \in \mathfrak{I}_K$, $BO_K = \{\alpha^{-1}xy : x \in I, y \in \mathcal{O}_K\} = \{\alpha^{-1}z : z \in I\} = B$.

We will now prove an important theorem.

**Theorem 5.5.1.** The set of fractional ideals $\mathfrak{I}_K$ is an abelian group under this multiplication operation.

**Proof.** We already proved closure under the operation and existence of identity, $\mathcal{O}_K$. Hence we only need to establish existence of inverses. For each ideal $I \subseteq \mathcal{O}_K$, define

$$I' = \{x \in K : xI \subseteq \mathcal{O}_K\},$$

and for each fractional ideal $B = \alpha^{-1}I \in \mathfrak{I}_K$, define $B' = \alpha I'$. Then each such $B'$ is again a fractional ideal, i.e. $B' \in \mathfrak{I}_K$ for every $B \in \mathfrak{I}_K$ (Problem 5.26). It is easy to see that $\mathcal{O}_K \subseteq B'$ for each $B' \in \mathfrak{I}_K$. Further, $\mathcal{O}'_K = \mathcal{O}_K$ (Problem 5.25). We will now prove that $B'$ is the inverse of $B$ for every $B \in \mathfrak{I}_K$. This is a lengthy proof, which will consist of a number of steps.

**Step 1.** For an ideal $I \subseteq \mathcal{O}_K$ we know that $\mathcal{O}_K \subseteq I'$. Let us prove that if $I$ is a proper ideal, then $\mathcal{O}_K \neq I'$. Let $M \subseteq \mathcal{O}_K$ be a maximal ideal such that $I \subseteq M$, then $M' \subseteq I'$. It will then suffice to prove that $M' \neq \mathcal{O}_K$, i.e. we want to find an element of $M'$ which is not in $\mathcal{O}_K$. First we need an auxiliary lemma.

**Lemma 5.5.2.** For every nonzero ideal $I \subseteq \mathcal{O}_K$, there exist prime ideals $P_1, \ldots, P_r \subseteq \mathcal{O}_K$ such that their product $P_1 \cdots P_r$ is contained in $I$.

**Proof.** Suppose not, then let $\mathcal{A}$ be the collection of all ideals in $\mathcal{O}_K$ for which this is not true, and let $I$ be a maximal element in $\mathcal{A}$ (since $\mathcal{O}_K$ is noetherian, such $I$ must exist). Then $I$ itself cannot be prime, and so exist ideals $J_1, J_2 \subseteq \mathcal{O}_K$ such that $J_1J_2 \subseteq I$ while $J_1, J_2$ are not in $I$. Let

$$A_1 = I + J_1, \ A_2 = I + J_2,$$

then $I \subseteq A_1, A_2$, while

$$A_1A_2 = (I + J_1)(I + J_2) = I^2 + I(J_1 + J_2) + J_1J_2 \subseteq I.$$
By maximality of $I$ in $\mathcal{A}$, it must be true that $A_1, A_2 \notin \mathcal{A}$, and hence there exist prime ideals $P_1, \ldots, P_r, Q_1, \ldots, Q_s$ in $\mathcal{O}_K$ such that

$$P_1 \cdots P_r \subseteq A_1, \ Q_1 \cdots Q_s \subseteq A_2,$$

but then

$$P_1 \cdots P_r, Q_1 \cdots Q_s \subseteq A_1 A_2 \subseteq I,$$

contradicting the choice of $I$. $\Box$

Back to our proof, let $0 \neq a \in M$, then the principal ideal $(a) \subseteq M$. Let $r$ be the smallest integer for which exists a collection of prime ideals $P_1, \ldots, P_r \in \mathcal{O}_K$ with the product $P_1 \cdots P_r \subseteq (a)$. Since $M$ is maximal (hence prime) at least one of $P_1, \ldots, P_r$ must be in $M$, say it is $P_1$. But in $\mathcal{O}_K$ every prime ideal is maximal, and hence $P_1 = M$. On the other hand, by minimality of $r$,

$$P_2 \cdots P_r \not\subseteq (a),$$

meaning that there exists some $b \in P_2 \cdots P_r$ which is not in $(a)$. This being said, $bP_1 = bM \subseteq (a)$, i.e. $ba^{-1}M \subseteq \mathcal{O}_K$, hence $ba^{-1} \in M'$. But since $b \notin (a) = a\mathcal{O}_K$, we have that $ba^{-1} \notin \mathcal{O}_K$, and thus we prove that $M' \neq \mathcal{O}_K$.

Step 2. Next, let $I \subseteq \mathcal{O}_K$ be an ideal and $T \subseteq K$ be a set such that $TI \subseteq I$. We will prove that $T \subseteq \mathcal{O}_K$. In other words, given $\theta \in T$, we will show that $\theta \in \mathcal{O}_K$. We know that $I$ is finitely generated as an ideal and is also a lattice of finite rank. Let $a_1, \ldots, a_n$ be a generating set for $I$, which is also an integral basis. In other words, for every $x \in I$, there exist $b_1, \ldots, b_n \in \mathbb{Z}$ such that

$$x = b_1a_1 + \cdots + b_na_n.$$

Since $\theta I \subseteq I$, we have

$$a_k\theta = b_{k1}a_1 + \cdots + b_{kn}a_n$$

for each $1 \leq k \leq n$, where $b_{kj} \in \mathbb{Z}$. In matrix form, this system of equations can be written as

$$Ba = \theta a,$$

where $B = (b_{kj})_{1 \leq k, j \leq n}$ is an integer matrix and $a = (a_1, \ldots, a_n)\top$ is a vector with coordinates in $I$. Then $\theta$ is an eigenvalue of $B$, which is a root of the monic polynomial

$$\det \begin{pmatrix} b_{11} - x & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} - x \end{pmatrix}$$

with integer coefficients. Hence $\theta$ is an algebraic integer in $K$, i.e. $\theta \in \mathcal{O}_K$. Thus $T \subseteq \mathcal{O}_K$.

Step 3. We are now ready to prove that $MM' = \mathcal{O}_K$ for a maximal ideal $M$ in $\mathcal{O}_K$. Indeed, it is clear that $MM' \subseteq \mathcal{O}_K$ is an ideal (as is $I$ for any ideal $I$ in $\mathcal{O}_K$), and

$$M \subseteq MM' \subseteq \mathcal{O}_K.$$

Since $M$ is maximal, then either $M = MM'$ or $MM' = \mathcal{O}_K$. Assume $M = MM'$, then by Step 2 we know that $M' \subseteq \mathcal{O}_K$, which contradicts Step 1. Hence we must have $MM' = \mathcal{O}_K$. 


Step 4. We now prove that \( II' = O_K \) for any nonzero ideal \( I \subseteq O_K \). Suppose not, then \( II' \not\subseteq O_K \). In fact, let us take \( I \) to be a maximal element in the set of all ideals \( J \) in \( O_K \) such that \( JJ' \subseteq O_K \). Let \( M \) be a maximal ideal in \( O_K \) containing \( I \). Then we know that 
\[
O_K \subseteq M' \subseteq I',
\]
hence
\[
I = IO_K \subseteq IM' \subseteq II' \not\subseteq O_K.
\]
Since \( IM' \) is an ideal in \( O_K \), the maximality condition on \( I \) implies that 
\[
(IM')(IM')' = I(M'(IM')') = O_K.
\]
This means that \( M'(IM')' \subseteq I' \), hence 
\[
O_K = I(M'(IM')') \subseteq II' \not\subseteq O_K,
\]
which is a contradiction. Hence \( II' = O_K \).

Step 5. It now easily follows that \( BB' = O_K \) for every \( B \in \mathfrak{F}_K \). Indeed, let \( B = \alpha^{-1}I \) be a fractional ideal in \( K \). Then \( B' = \alpha I' \), and so 
\[
BB' = (\alpha^{-1}I)(\alpha I') = II' = O_K.
\]
This completes the argument, showing that \( \mathfrak{F}_K \) is the abelian group of fractional ideals in \( K \). \( \square \)

This result and its proof have some important implications, the first of which is a certain weaker analogue of unique factorization in \( O_K \).

**Theorem 5.5.3.** Every nonzero ideal in \( O_K \) can be written uniquely (up to permutation) as a product of prime ideals.

**Proof.** We first show existence of such a factorization, and then its uniqueness. Suppose such a factorization does not exist for every ideal, and let \( S \) be the set of all ideals in \( O_K \) which do not have such a factorization. Since \( O_K \) is noetherian, \( S \) has a maximal element, call it \( I \). Then \( I \) itself is not a prime ideal, hence not a maximal ideal in \( O_K \). Let \( M \) be a maximal ideal in \( O_K \) such that \( I \subseteq M \). Then by the argument above (Step 4 in the proof of Theorem 5.5.1), 
\[
I \subseteq IM' \subseteq O_K.
\]
Since \( I \) is maximal without a prime factorization, the ideal \( IM' \) must have a factorization into prime ideals, say 
\[
IM' = P_1 \cdots P_r
\]
for some prime ideals \( P_1, \ldots, P_r \subseteq O_K \). Then, since \( M'M = O_K \), 
\[
P_1 \cdots P_r M = (IM')M = I(M'M) = I,
\]
which is a factorization of \( I \) into a product of prime ideals.

We now establish uniqueness of such factorization. Let \( r \) be the length of a shortest factorization into prime ideals for an ideal \( I \subseteq O_K \). We argue by induction on \( r \). If \( r = 1 \), then \( I \) itself is prime, and we are done. Assume the result is true for all ideals with length of prime factorization \( \leq r - 1 \). Let us prove it for \( r \). Suppose 
\[
I = P_1 \cdots P_r = Q_1 \cdots Q_s
\]
are two factorizations of \( I \) into primes, \( s \geq r \). Since the ideals \( P_1, \ldots, P_r, Q_1, \ldots, Q_s \) are prime (hence maximal), we must have \( P_i = Q_i \) for some \( 1 \leq i \leq s \); in fact,
after rearranging, if necessary, we can assume \( i = 1 \). Multiplying both sides of this equality by \( P'_1 \), we obtain
\[
P_2 \cdots P_r = Q_2 \cdots Q_s,
\]
since \( P'_1 P_1 = Q'_1 Q_1 = \mathcal{O}_K \). By induction hypothesis, this factorization is unique, and hence we are done. \( \square \)
5.6. Further properties of ideals

We continue studying structure of ideals here with a view towards factorization of elements into irreducibles in rings of algebraic integers of number fields. The main goal of this section is to give a vital condition for such a factorization to be unique. We start with an important property of the norm of an ideal: it is a multiplicative function, same as norm of an element.

**Lemma 5.6.1.** Let \( K \) be a number field, and \( I, J \) be nonzero ideals in the ring \( \mathcal{O}_K \).

Then
\[
\mathbb{N}_K(IJ) = \mathbb{N}_K(I)\mathbb{N}_K(J).
\]

**Proof.** It is sufficient to prove this statement in the case when one of these ideals, say \( J \), is prime, let us call it \( P \) (Problem 5.27). In other words, we want to prove that
\[
|\mathcal{O}_K/IP| = |\mathcal{O}_K/I| \cdot |\mathcal{O}_K/P|.
\]

First let us prove that
\[
|\mathcal{O}_K/IP| = |\mathcal{O}_K/I| \cdot |I/IP|.
\]
For this, let us define a map \( \phi: \mathcal{O}_K/IP \to \mathcal{O}_K/I \), given by \( \phi(x+IP) = x + I \). This is a surjective ring homomorphism (Problem 5.28). By First Isomorphism Theorem for groups,
\[
(\mathcal{O}_K/IP) / \text{Ker}(\phi) \cong \phi(\mathcal{O}_K/IP) = \mathcal{O}_K/I,
\]
and since these groups are finite, we have
\[
|\mathcal{O}_K/IP| = |\text{Ker}(\phi)| \cdot |\mathcal{O}_K/I|.
\]
Notice that \( \text{Ker}(\phi) \) consists of all cosets of \( IP \) which are in \( I \), i.e. \( \text{Ker}(\phi) = I/IP \). This establishes (5.11).

Next we prove that
\[
|I/IP| = |\mathcal{O}_K/P|.
\]
Let \( B \subseteq \mathcal{O}_K \) be an ideal such that
\[
IP \subseteq B \subseteq I.
\]
Let \( I' \) be the inverse of \( I \) in the group of \( \mathfrak{F}_K \) of fractional ideals of \( K \), then
\[
I'(IP) = P \subseteq IB \subseteq I' = \mathcal{O}_K,
\]
and since \( P \) is a prime ideal in \( \mathcal{O}_K \) (hence maximal), we must have \( I'B \) either equal to \( P \) or to \( \mathcal{O}_K \). This means that \( B \) is either equal to \( IP \) or \( I \), i.e. there is no ideal between \( IP \) and \( I \). Let \( \alpha \in I \setminus IP \), then we must have
\[
IP + \langle \alpha \rangle = I.
\]
Define a function \( f_\alpha: \mathcal{O}_K \to I/IP \) by \( f_\alpha(x) = \alpha x + IP \). This is an abelian group homomorphism, which is surjective:
\[
f_\alpha(\mathcal{O}_K) = (\alpha \mathcal{O}_K + IP)/IP = (IP + \langle \alpha \rangle)/IP = I/IP.
\]
Therefore \( \mathcal{O}_K/\text{Ker}(f_\alpha) \cong I/IP \). Clearly, \( \text{Ker}(f_\alpha) \neq \mathcal{O}_K \), since \( I \neq IP \). On the other hand, for every \( x \in P \),
\[
f_\alpha(x) = \alpha x + IP = 0
\]
in $I/IP$, since $\alpha x \in IP$ ($\alpha$ is in $I$ and $x$ is in $P$). Therefore $P \subseteq \text{Ker}(f_\alpha)$, and since $P$ is a maximal ideal, we have $P = \text{Ker}(f_\alpha)$. Therefore $O_K/P \cong I/IP$, and (5.12) follows. Combining (5.11) with (5.12) proves (5.10).

We next look closer at the norm values of ideals.

**Lemma 5.6.2.** Let $I \subseteq O_K$ be a nonzero ideal. Then:

1. If $N_K(I)$ is a prime in $\mathbb{Z}$, then $I$ is a prime ideal in $O_K$.
2. $N_K(I) \in I$.
3. If $I \subset O_K$ is a prime ideal, then $N_K(I) = p^m$ for some prime $p \in \mathbb{Z}$ and $m \leq d = [K : \mathbb{Q}]$.

**Proof.** To prove part (1), suppose that $I = AB$ for some two ideals $A, B \subseteq O_K$. Then, by multiplicativity of the norm,

$$N_K(I) = N_K(A)N_K(B) = p \in \mathbb{Z},$$

where $p$ is a prime number. Hence either $N_K(A)$ or $N_K(B)$ must be equal to 1, say it is $N_K(A)$. This means that $|O_K/A| = 1$, and so $A = O_K$, hence

$$I = AB = O_KB = B.$$ Therefore $I$ does not have any nontrivial factorization, hence it is a prime ideal.

To prove part (2), let $x + I \in O_K/I$. Since $O_K/I$ is an additive abelian group of order $N_K(I)$, we must have

$$N_K(I)(x + I) = N_K(I)x = I,$$

meaning that $N_K(I)x \in I$ for every $x \in O_K$. Then take $x = 1$, and we see that $N_K(I) \in I$.

To prove part (3), assume that

$$N_K(I) = p_1^{m_1}\cdots p_r^{m_r}$$

for some rational primes $p_1, \ldots, p_r \in \mathbb{Z}$ and positive integers $m_1, \ldots, m_r$. By part (2), we have

$$p_1^{m_1}\cdots p_r^{m_r} \in I.$$ Since $I$ is a primed ideal, we must have $p_i \in I$ for some $1 \leq i \leq r$. Suppose $q \in I$ is a rational integer prime different from $p_i$, then, by Euclid’s Division Lemma, there exist some $a, b \in \mathbb{Z}$ such that

$$1 = ap_i + bq,$$

and so $1 \in I$, meaning that $I = O_K$. However, a prime ideal has to be proper. Thus $p_i$ is the only rational integer prime contained in $I$, which means that $I \mid p_iO_K$ and $I \nmid qO_K$ for any prime $q \neq p_i$. Therefore

$$N_K(I)|N_K(p_i) = N_K(p_i) = \prod_{j=1}^d \sigma_j(p_i) = p_i^d,$$

where $\sigma_1, \ldots, \sigma_d$ are embeddings of $K$. Therefore $N_K(I) = p_i^m$ for some $m \leq d$. □

The next lemma contains some key finiteness observations about ideals of given norm.

**Lemma 5.6.3.** Let $K$ be a number field with the ring of integers $O_K$. The following are true:
Let $I \subseteq \mathcal{O}_K$ be an ideal. There exist only finitely many ideals $J \subseteq \mathcal{O}_K$ such that $J \mid I$. Let $m \in \mathbb{Z}$, $m \neq 0$. There exist only finitely many ideals $I \subseteq \mathcal{O}_K$ such that $m \in I$. Let $m \in \mathbb{Z}_{>0}$. There exist only finitely many ideals $I \subseteq \mathcal{O}_K$ of norm $m$.

**Proof.** To prove part (1), we use Theorem 5.5.3: since there is a unique factorization $I$ into prime ideals, all ideals dividing $I$ must be products of some subcollections of these primes ideals, hence there can only be finitely many of them.

For part (2), let $m \in \mathbb{Z}$ be nonzero and let $I \subseteq \mathcal{O}_K$ be an ideal such that $m \in I$. Then $I \mid m\mathcal{O}_K$, but by part (1) the ideal $m\mathcal{O}_K$ can have only finitely many divisors. Thus $m$ can belong to only finitely many ideals in $\mathcal{O}_K$.

Finally, for part (3) let $I$ be an ideal of norm $m$, then by part (2) of Lemma 5.6.2 above, $m \in I$. However, by part (2), there can be only finitely many ideals $I$ so that $m \in I$. Thus there can be only finitely many ideals of norm $m$. □

We need one more technical lemma before we can prove the main theorem of this section.

**Lemma 5.6.4.** Let $I, J \subseteq \mathcal{O}_K$ be nonzero ideals. Then there exists an element $\alpha \in I$ such that

$$\alpha I' + J = \mathcal{O}_K,$$

where $I'$ is the inverse of $I$ in $\mathfrak{f}_K$.

**Proof.** Let $\alpha$ be any element of $I$, then $\alpha I'$ is an ideal in $\mathcal{O}_K$, since

$$I' = \{x \in K : xI \subseteq \mathcal{O}_K\}.$$ 

Thus $\alpha I' + J$ is an ideal in $\mathcal{O}_K$, which contains ideals $\alpha I'$ and $J$: in fact, it is the smallest such ideal (with respect to inclusion). This means that $\alpha I' + J$ is the greatest common divisor of $\alpha I'$ and $J$. Let

$$J = P_1 \cdots P_r$$

be the unique factorization of $J$ into prime ideals. Then

$$\alpha I' + J \mid J = P_1 \cdots P_r,$$

and $\alpha I' + J \subseteq \alpha I' + P_i$ for each $1 \leq i \leq r$, since $J \subseteq P_i$ for each $i$. Hence

$$\alpha I' + J = \bigcap_{i=1}^r (\alpha I' + P_i).$$

We will now construct and $\alpha \in I$ such that $\alpha I' + P_i = \mathcal{O}_K$ for each $1 \leq i \leq r$. Since each $P_i$ is a maximal ideal, it is sufficient to construct $\alpha \in I$ such that $\alpha I' \nsubseteq P_i$ for all $1 \leq i \leq r$; if $\alpha I' \subseteq P_i$, then

$$P_i \not\subseteq \alpha I' + P_i \subseteq \mathcal{O}_K,$$

which by maximality of $P_i$ means that $\alpha I' + P_i = \mathcal{O}_K$. Notice that

$$\alpha I' \nsubseteq P_i \iff P_i \nmid \alpha I' \iff IP_i \nmid \alpha \mathcal{O}_K \iff \alpha \notin IP_i$$

for all $1 \leq i \leq r$. Hence we need to construct an element $\alpha \in I \setminus (\bigcup_{i=1}^r IP_i)$.

Notice that for each $1 \leq i \leq r$, $IP_i \nsubseteq I$. Thus if $r = 1$, we can take any element $\alpha \in I \setminus IP_1$. Assume $r > 1$, and for each $1 \leq m \leq r$, define

$$I_m = IP_1 \cdots P_{m-1}P_{m+1} \cdots P_r,$$
then $IP_m = IJ$. For each $m$, pick an element $\alpha_m \in I_m \setminus IJ$, and define
\[
\alpha = \alpha_1 + \cdots + \alpha_r.
\]
Then $\alpha \in I$, since each $\alpha_m \in I_m \subseteq I$. Assume that $\alpha \in IP_m$ for some $1 \leq m \leq r$. Notice that for each $j \neq m$, $\alpha_j \in I_j \subseteq IP_m$, and so
\[
\alpha_m = \alpha - (\alpha_1 + \cdots + \alpha_{m-1} + \alpha_{m+1} + \cdots + \alpha_r) \in IP_m.
\]
This contradicts our choice of $\alpha_m$, meaning that $\alpha \notin IP_m$ for any $1 \leq m \leq r$. This completes the proof. □

We are now ready for the main result of this section. Recall that an integral domain is called principal (abbreviated PID) if every ideal in it can be generated by one element.

**Theorem 5.6.5.** Let $K$ be a number field and $\mathcal{O}_K$ its ring of integers. The $\mathcal{O}_K$ is a UFD if and only if it is a PID.

**Proof.** Every PID is a UFD (Problem 5.29: this is a standard theorem, found in any algebra book), so we will only prove the reverse implication. Suppose $\mathcal{O}_K$ is a UFD. Let $I \subseteq \mathcal{O}_K$ be an ideal and let
\[
I = P_1 \cdots P_r
\]
be its factorization into prime ideals. If each $P_i$ is principal, say $P_i = x_i \mathcal{O}_K$, then $I = (x_1 \cdots x_r)\mathcal{O}_K$ is also principal. Hence we only need to prove that primes ideals in $\mathcal{O}_K$ are principal. Let $P \subset \mathcal{O}_K$ be a prime ideal and let $m = N_K(P)$. Then Lemma 5.6.2 guarantees that $m \in P$, i.e. $P \mid m\mathcal{O}_K$. Let us write $m = x_1 \cdots x_k$ be the factorization of $m$ into irreducibles in $\mathcal{O}_K$. Since $\mathcal{O}_K$ is a UFD, we know that irreducibles are primes, and so the principal ideals $\langle x_1 \rangle, \ldots, \langle x_k \rangle$ are prime ideals in $\mathcal{O}_K$. Then we have
\[
P \mid m\mathcal{O}_K = \langle x_1 \rangle, \ldots, \langle x_k \rangle,
\]
which means that $P \mid \langle x_i \rangle$ for some $1 \leq i \leq k$, since $P$ is prime. Since prime ideals are maximal in $\mathcal{O}_K$, we must have $P = \langle x_i \rangle$, so $P$ is a principal ideal. This completes the proof. □

Finally, we record here a simple, but somewhat surprising corollary of Lemma 5.6.4. Many (in some sense, most) rings $\mathcal{O}_K$ do not have unique factorization into irreducibles, and hence have non-principal ideals by Theorem 5.6.5. It turns out, however, that every ideal in $\mathcal{O}_K$ can be generated by at most two elements.

**Corollary 5.6.6.** Let $I \subseteq \mathcal{O}_K$ be a nonzero ideal and let $\beta$ be a nonzero element of $I$. Then there exists an element $\alpha \in I$ such that $I$ is generated by the pair $\alpha, \beta$.

**Proof.** Let $J = \beta I'$, and ideal in $\mathcal{O}_K$. By Lemma 5.6.4 there exists $\alpha \in I$ such that $\alpha I' + J = \mathcal{O}_K$, i.e.
\[
\alpha I' + \beta I' = \mathcal{O}_K.
\]
Multiplying both sides of this equality by $I$, we obtain:
\[
\alpha I' I + \beta I' I = \alpha \mathcal{O}_K + \beta \mathcal{O}_K = I\mathcal{O}_K = I.
\]
Thus $I$ is generated by $\alpha, \beta$. □
5.7. Minkowski embedding

We have previously proved that any nonzero ideal in the ring of integers \( \mathcal{O}_K \) of a number field \( K \) is a lattice, i.e., a free abelian group. In fact, ideals can be embedded into a Euclidean space and viewed as lattices there, which helps to study their properties. To do this, we use embeddings of our number field to form the important Minkowski embedding.

As usual, let \( K \) be a number field of degree \( d \) over \( \mathbb{Q} \), and let \( \sigma_1, \ldots, \sigma_d : K \rightarrow \mathbb{C} \) be its embeddings. We will distinguish between real and complex embeddings: \( \sigma_i \) is said to be real if the field \( \sigma_i(K) \) is contained in \( \mathbb{R} \), and complex otherwise. Notice that complex embeddings come in conjugate pairs: if \( \sigma_i \) is complex, then there is its conjugate embedding \( \overline{\sigma_i} \) given by

\[
\overline{\sigma_i}(x) := \overline{\sigma_i(x)}
\]

for every \( x \in K \). Let us order the embeddings

\[
\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \overline{\sigma_{r+1}}, \ldots, \sigma_{r+s}, \overline{\sigma_{r+s}},
\]

where \( \sigma_1, \ldots, \sigma_r \) are real and \( \sigma_{r+1}, \overline{\sigma_{r+1}}, \ldots, \sigma_{r+s}, \overline{\sigma_{r+s}} \) are complex. Then \( d = r + 2s \), and we can define a map

\[
\Sigma := (\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \ldots, \sigma_{r+s}) : K \rightarrow \mathbb{R}^r \times \mathbb{C}^s,
\]

given by \( \Sigma(x) = (\sigma_1(x), \ldots, \sigma_i(x), \sigma_{r+1}(x), \ldots, \sigma_{r+s}(x)) \) for every \( x \in K \). We can identify \( \mathbb{C}^s \) with \( \mathbb{R}^{2s} \), thinking of \( \Sigma(x) \) as

\[
(\sigma_1(x), \ldots, \sigma_r(x), \Re(\sigma_{r+1}(x)), \Im(\sigma_{r+1}(x)), \ldots, \Re(\sigma_{r+s}(x)), \Im(\sigma_{r+s}(x))).
\]

Let us now consider images of ideals in \( \mathcal{O}_K \) under this Minkowski embedding of the number field into \( \mathbb{R}^d \).

**Lemma 5.7.1.** Let \( M \) be a lattice contained in \( K \) and \( x_1, \ldots, x_d \) be a \( \mathbb{Z} \)-basis for \( M \). Then \( \Sigma(M) \) is a lattice of rank \( d \) in \( \mathbb{R}^d \) with determinant

\[
\det(\Sigma(M)) = 2^{-s} |\det(A) C|,
\]

where \( A \) is an

\[
A = (\sigma_i(x_j))_{1 \leq i \leq r, 1 \leq j \leq d}, \quad C = (\sigma_{r+i}(x_j) \overline{\sigma_{r+i}(x_j)})_{1 \leq i \leq s, 1 \leq j \leq d}.
\]

**Proof.** Notice that \( \det(\Sigma(M)) \) is equal to the absolute value of the determinant of the block matrix \( (A B) \), where \( A \) is the \( d \times r \) matrix defined in (5.14) and \( B \) is the \( d \times 2s \)

\[
B = (\Re(\sigma_{r+i}(x_j)) \Im(\sigma_{r+i}(x_j)))_{1 \leq i \leq s, 1 \leq j \leq d}
\]

For any complex number \( z \),

\[
\Re(z) = \frac{1}{2}(z + \overline{z}), \quad \Im(z) = \frac{1}{2i}(z - \overline{z}).
\]

With this in mind, it is easy to see that the matrix \( B \) is column-equivalent to the matrix

\[
\left( \frac{1}{2} (\sigma_{r+i}(x_j) + \overline{\sigma_{r+i}(x_j)}) \frac{1}{2i} (\sigma_{r+i}(x_j) - \overline{\sigma_{r+i}(x_j)}) \right)_{1 \leq i \leq s, 1 \leq j \leq d}.
\]

This means that

\[
\det(\Sigma(M)) = \left| \left( \frac{1}{2} \right)^{2s} \left( \frac{1}{i} \right)^s \det(A B') \right| = 2^{-2s} |\det(A) B'|,
\]

where \( B' \) is obtained from \( B \) by column permutations. This finishes the proof.
where
\[ B' = (\sigma_{r+i}(x_j) + \bar{\sigma}_{r+i}(x_j)) \ (\sigma_{r+i}(x_j) - \bar{\sigma}_{r+i}(x_j)) \}_{1 \leq i \leq s, 1 \leq j \leq d}, \]
a matrix column-equivalent to
\[ B'' = (2\sigma_{r+i}(x_j) \ \bar{\sigma}_{r+i}(x_j)) \}_{1 \leq i \leq s, 1 \leq j \leq d}. \]
Thus
\[ \det(\Sigma(M)) = 2^{-2s} |\det(A B'')| = 2^{-s} |\det(A C)|, \]
and so we have (5.13). Notice that this determinant cannot be equal to 0 (Problem 5.30), and hence \( \Sigma(M) \) is a lattice of full rank in \( \mathbb{R}^d \). \( \square \)

**Corollary 5.7.2.** If \( I \subseteq \mathcal{O}_K \) is an ideal, then
\[ \det(\Sigma(I)) = 2^{-s} |\Delta_K|^{1/2} N_K(I). \]
In particular,
\[ \det(\Sigma(\mathcal{O}_K)) = 2^{-s} |\Delta_K|^{1/2}. \]

**Proof.** By Lemma 5.4.5,
\[ N_K(I) = \left| \frac{\Delta(\alpha_1, \ldots, \alpha_d)}{\Delta_K} \right|^{1/2}, \]
where \( \alpha_1, \ldots, \alpha_d \) is a \( \mathbb{Z} \)-basis for \( I \). Then, by Lemma 5.7.1 (and in the notation of that lemma),
\[ \det(\Sigma(I)) = 2^{-s} |\det(A C)| = 2^{-s} |\Delta(\alpha_1, \ldots, \alpha_d)|^{1/2} = 2^{-s} N_K(I) |\Delta_K|^{1/2}. \]
If \( I = \mathcal{O}_K \), then \( N_K(I) = 1 \). \( \square \)

**Lemma 5.7.3.** Let \( d = r + 2s \) and let \( \Lambda \subset \mathbb{R}^d \) be a lattice of full rank with \( \det(\Lambda) = D \). Let \( c_1, \ldots, c_r, d_1, \ldots, d_s \in \mathbb{R}_{>0} \) be such that
\[ c_1 \cdots c_r d_1 \cdots d_s > \left( \frac{4}{\pi} \right)^s D. \]
Let
\[ X = \{ \mathbf{x} = (x_1, \ldots, x_r, y_{11}, y_{12}, \ldots, y_{s1}, y_{s2}) \in \mathbb{R}^d : x_i < c_i \ \forall \ 1 \leq i \leq r, \ y_{j1}^2 + y_{j2}^2 < d_j \ \forall \ 1 \leq j \leq s \}. \]
Then there exists \( \mathbf{0} \neq \mathbf{x} \in X \cap \Lambda \).

**Proof.** This lemma is proved by an application of Minkowski’s Convex Body Theorem. First notice that the set \( X \) is a Cartesian product of intervals \([-c_i, c_i]\) for all \( 1 \leq i \leq r \) and circles of radius \( \sqrt{d_j} \) for \( 1 \leq j \leq s \). Therefore \( X \) is convex \( \mathbf{0} \)-symmetric and its volume is
\[ \text{Vol}(X) = (2c_1) \cdots (2c_r) \cdot (\pi d_1) \cdots (\pi d_s) > 2^r \pi^s \left( \frac{4}{\pi} \right)^s D = 2^d D. \]
Then Theorem 1.3.2 (see also Problem 1.13) guarantees that there exists a nonzero point \( \mathbf{x} \in X \cap \Lambda \). \( \square \)

We can now apply the lemma above to prove that every ideal has a nonzero element of small norm.
Corollary 5.7.4. Let $I \subseteq \mathcal{O}_K$ be a nonzero ideal. There exists a nonzero element $\alpha \in I$ such that

$$N_K(\alpha) \leq \left(\frac{2}{\pi}\right)^s N_K(I) |\Delta_K|^{1/2},$$

where $s$ is the number of conjugate pairs of complex embeddings of the number field $K$, as above.

Proof. Let $r$ be the number of real embeddings of $K$, $\varepsilon > 0$ and let

$$c_1, \ldots, c_r, d_1, \ldots, d_s \in \mathbb{R}_{>0}$$

be such that

$$c_1 \cdots c_r \cdot d_1 \cdots d_s = \left(\frac{2}{\pi}\right)^s N_K(I) |\Delta_K|^{1/2} + \varepsilon.$$ 

Let $\Lambda = \Sigma(I)$ and let $X = X_\varepsilon$ be as in Lemma 5.7.3. Applying Corollary 5.7.2, we see that

$$c_1 \cdots c_r \cdot d_1 \cdots d_s > \left(\frac{4}{\pi}\right)^s \det(\Lambda).$$ 

Therefore Lemma 5.7.3 implies that there exists a nonzero point in $X_\varepsilon \cap \sigma(I)$, and hence this point is of the form $\Sigma(\alpha)$ for some $\alpha \in I$. We can then compute

$$N_K(\alpha) = \prod_{i=1}^r |\sigma_i(\alpha)| \times \prod_{j=1}^s |\sigma_{r+j}(\alpha)\bar{\sigma}_{r+j}(\alpha)| = \prod_{i=1}^r |\sigma_i(\alpha)| \times \prod_{j=1}^s |\sigma_{r+j}(\alpha)|^2$$

$$< c_1 \cdots c_r \cdot d_1 \cdots d_s = \left(\frac{2}{\pi}\right)^s N_K(I) |\Delta_K|^{1/2} + \varepsilon.$$ 

Since $\Lambda$ is discrete in $\mathbb{R}^d$, there are only finitely such $\alpha$ for every $\varepsilon > 0$. Hence as $\varepsilon \to 0$, the intersection of all sets

$$\{\alpha \in I : \Sigma(\alpha) \in X_\varepsilon\}$$

will be nonempty, i.e. there exists a nonzero $\alpha \in I$ satisfying the bound of the lemma. \qed
5.8. The class group

Now that we understand that non-uniqueness of factorization into irreducibles in a ring $\mathcal{O}_K$ is equivalent to existence of non-principal ideals, we look more closely at the structure of the group of fractional ideals $\mathfrak{F}_K$. We say that a fractional ideal $B \in \mathfrak{F}_K$ is principal if $B = \alpha^{-1}I$ for $\alpha \in \mathcal{O}_K$ and $I \subseteq \mathcal{O}_K$ a principal ideal. The set $\mathfrak{P}_K$ of all principal fractional ideals is then a subgroup of $\mathfrak{F}_K$ (Problem 5.31): in fact, since $\mathfrak{F}_K$ is an abelian group, the subgroup $\mathfrak{P}_K$ is normal.

**Definition 5.8.1.** The ideal class group, or simply the class group of the number field $K$ is the quotient group $\text{Cl}(K) := \mathfrak{F}_K / \mathfrak{P}_K$.

Elements of this group are called ideal classes, and the order $h_K := |\text{Cl}(K)|$ of the class group is called the class number of the number field $K$.

Two fractional ideals $B_1$ and $B_2$ are in the same ideal class, denoted by $B_1 \sim B_2$ if and only if there exists some $a, b \in \mathcal{O}_K$ such that $\langle a \rangle B_1 = \langle b \rangle B_2$; this is an equivalence relation on $\mathfrak{F}_K$ (Problem 5.32). We immediately have the following important consequence of Theorem 5.6.5.

**Corollary 5.8.1.** The ring of integers $\mathcal{O}_K$ of the number field $K$ is a UFD if and only if the class number $h_K = 1$.

**Proof.** Theorem 5.6.5 asserts that $\mathcal{O}_K$ is a UFD if and only if any ideal $I \subseteq \mathcal{O}_K$ is principal. This is equivalent to saying that every fractional ideal $B \in \mathfrak{F}_K$ is principal, since $B = \alpha^{-1}I$ for some $\alpha \in \mathcal{O}_K$ and $I \subseteq \mathcal{O}_K$ an ideal. This, in turn, is equivalent to $\mathfrak{P}_K$ being all of $\mathfrak{F}_K$, i.e. the class group $\text{Cl}(K)$ being trivial and hence having order 1. $\square$

Hence we have a nice quantitative test to check if an analogue of the Fundamental Theorem of Arithmetic holds in $\mathcal{O}_K$: compute the class number $h_K$ and check if it is equal to 1. The problem is that $h_K$ can be very hard to compute. In fact, it is not even clear whether it is finite. The truth is, it is, however proving it requires all the machinery we developed thus far. The finiteness of the class number, which we are about to establish, is one of the greatest achievements of the classical Algebraic Number Theory: it was first proved by Minkowski with the use of his Geometry of Numbers. In fact, discovery of the kingdom referred to in the epigraph to this text alludes precisely to this result.

**Theorem 5.8.2.** The class number $h_K$ of any number field $K$ is finite.

To prove this theorem, we first need an auxiliary lemma, which follows from our results of Section 5.7.

**Lemma 5.8.3.** Every ideal class in $\text{Cl}(K)$ contains an ideal $I \subseteq \mathcal{O}_K$ with

$$N_K(I) \leq \left(\frac{2}{\pi}\right)^{s} |\Delta_K|^{1/2}.$$ (5.15)

**Proof.** Since every ideal class contains an ideal (Problem 5.33), we only need to prove that every ideal in $\mathcal{O}_K$ is equivalent to some ideal $I \subseteq \mathcal{O}_K$ satisfying (5.15). Let $J \subseteq \mathcal{O}_K$ be an ideal and let $M$ be an ideal equivalent to $J'$; then

$$JM \sim JJ' \sim \mathcal{O}_K.$$
By Corollary 5.7.4, there exists a nonzero element \( \alpha \in M \) such that
\[
|N_K(\alpha)| \leq \left( \frac{2}{\pi} \right)^s N_K(M) |\Delta_K|^{1/2}.
\]
Since \( \alpha \in M \), it must be true that \( M \mid \langle \alpha \rangle \), i.e. there exists some ideal \( I \subseteq \mathcal{O}_K \) such that \( \langle \alpha \rangle = IM \). Then, by multiplicativity of the norm,
\[
N_K(I)N_K(M) = N_K(\langle \alpha \rangle) = |N_K(\alpha)| \leq \left( \frac{2}{\pi} \right)^s N_K(M) |\Delta_K|^{1/2},
\]
which means that
\[
N_K(I) \leq \left( \frac{2}{\pi} \right)^s |\Delta_K|^{1/2}.
\]
Now,
\[
IM = \langle \alpha \rangle \in \mathfrak{P}_K,
\]
which means that \( IM \sim \mathcal{O}_K \), and so \( I \sim M' \sim (J')' = J \). Hence the ideal class of \( J \) contains an ideal with norm bounded as in (5.15). This completes the proof. \( \square \)

**Proof of Theorem 5.8.2.** By Lemma 5.8.3 we know that every ideal class in \( \text{Cl}(K) \) must contain an ideal of norm bounded as in (5.15). This means that the \( h_K \), the number of ideal classes has to be no bigger than the number of ideals \( I \subseteq \mathcal{O}_K \) with norm less or equal than \( \left( \frac{2}{\pi} \right)^s |\Delta_K|^{1/2} \), a finite positive number. Now Lemma 5.6.3 readily implies that there can be only finitely many such ideals. This completes the proof. \( \square \)

Theorem 5.8.2 establishes the finiteness of the class number \( h_K \), but does not provide a direct way to compute it. In fact, explicit computation of the class number for a given number field \( K \) can be very difficult; while there are some known class number formulas, they are usually in terms of other invariants of the number field that are also quite hard to compute. In particular, the classification of number fields with class number equal to 1, i.e. those whose rings of integers allow unique factorization into irreducibles is far from complete even in the case of degree 2. Quadratic number fields are of the form \( K = \mathbb{Q}(\sqrt{D}) \) where \( D \) is a squarefree integer: \( K \) is called real if \( D > 0 \) and imaginary if \( D < 0 \). The problem of determining the class number of quadratic number fields goes back to Gauss, who stated several highly influential conjectures. For imaginary quadratics, Gauss conjectured that \( h_{\mathbb{Q}(\sqrt{D})} \to \infty \) as \( D \to -\infty \): this was proved to be true by Heilbronn in 1934. Furthermore, in 1935 Siegel showed that \( h_{\mathbb{Q}(\sqrt{D})} \) grows approximately like \( \sqrt{|D|} \) as \( D \to -\infty \). Gauss also compiled lists of imaginary quadratic number fields of low class number, such as 1, 2, 3, believing them to be complete. The case \( h_K = 1 \), where Gauss listed 9 fields received especially a lot of attention. It was proved by Heilbronn and Linfoot in 1934 that there can be at most 10 such fields, and then Heegner in 1952 (and later independently Stark and Baker) produced the 10-th such field. As a result the full list of in integers \( D \) such that the imaginary quadratic \( \mathbb{Q}(\sqrt{D}) \) has class number 1 is:

\[-1, -2, -3, -7, -11, -19, -43, -67, -163.\]

Interestingly, \( K = \mathbb{Q}(\sqrt{-19}) \) is the first example of a number field with \( \mathcal{O}_K \) being a PID, but not Euclidean, meaning that there is no Euclidean algorithm possible on \( \mathcal{O}_K \). The full lists of imaginary quadratics with class numbers up to 100 have now
been completed (Watkins, 2004). To contrast, the situation is far more complicated with real quadratic fields: here the original Gauss conjecture that there are infinitely many of them with class number 1 remains open.
5.9. Problems

Problem 5.1. Suppose that \( L \) is a field extension of \( K \). Prove that \( L \) is \( K \)-vector space.

Problem 5.2. Let \( K \) and \( L \) be subfields of \( \mathbb{C} \). Prove that their intersection \( K \cap L \) is also a subfield of \( \mathbb{C} \). Use this fact to conclude uniqueness of the extension \( K(\alpha_1, \ldots, \alpha_n) \) for \( \alpha_1, \ldots, \alpha_n \in \mathbb{C} \), as defined above.

Problem 5.3. Let \( K \subseteq \mathbb{C} \) be a subfield, \( \alpha, \beta \in \mathbb{C} \), and let \( K_1 = K(\alpha) \), \( K_2 = K(\beta) \), \( L = K(\alpha, \beta) \). Prove that \( L = K_1(\beta) = K_2(\alpha) \). Conclude that
\[
[L : K] = [L : K_1][K_1 : K] = [L : K_2][K_2 : K].
\]

Problem 5.4. Prove that \( \dim_{\mathbb{Q}} \mathbb{Q}[\sqrt{2}] = 2 \).

Problem 5.5. Prove that \( K[\alpha] \subseteq K(\alpha) \) for any subfield \( K \subseteq \mathbb{C} \) and \( \alpha \in \mathbb{C} \).

Problem 5.6. Let \( K \subseteq \mathbb{C} \) be a finite algebraic extension of \( \mathbb{Q} \) and let \( \alpha \in \mathbb{C} \) be an algebraic number. Prove parts (3) and (4) of Theorem 5.1.1 with \( \mathbb{Q} \) replaced by \( K \).

Problem 5.7. Let \( K \subseteq \mathbb{C} \) be a field and \( \alpha \in \mathbb{C} \). Prove that \( K[\alpha] \subseteq \mathbb{C} \) is a ring under the operations on \( \mathbb{C} \), as is \( K[x] \) under the addition and multiplication of polynomials.

Problem 5.8. Let \( \alpha \in \mathbb{C} \), and define a map \( \varphi : K[x] \to K[\alpha] \) by
\[
\varphi(f(x)) = f(\alpha)
\]
for every \( f(x) \in K[x] \). Prove that \( \varphi \) is a ring homomorphism. Describe its kernel and specify under what conditions on \( \alpha \) is this an isomorphism.

Problem 5.9. Prove that \( \mathcal{O}_\mathbb{Q} = \mathbb{Z} \). Due to this property, elements of \( \mathbb{Z} \) are often called rational integers. Prove also that \( \mathbb{Z} \subseteq \mathcal{O}_K \) for any number field \( K \).

Problem 5.10. Let \( K \subseteq \mathbb{C} \) be a finite extension of \( \mathbb{Q} \). Without using the Primitive Element Theorem, prove that there must exist algebraic numbers \( \alpha_1, \ldots, \alpha_n \in \mathbb{C} \) such that \( K = \mathbb{Q}(\alpha_1, \ldots, \alpha_n) \).

Problem 5.11. Let \( \alpha \) be an algebraic integer of degree \( d \geq 1 \) and let \( n \) be a positive integer. Prove that the numbers
\[
1, \alpha, \ldots, \alpha^n
\]
are linearly independent over \( \mathbb{Z} \) if and only if \( n < d \).

Problem 5.12. Prove that this \( G \) is a subgroup of \( \mathbb{C} \) under the usual addition of complex numbers, and hence is an additive abelian group.
Problem 5.13. Let A and B be subrings of the same ring R. Prove that A ∩ B is also a ring. Use this fact to prove that for any number field K, the set \( \mathcal{O}_K \) of all algebraic integers in K is a commutative ring with identity.

Problem 5.14. Prove that each \( \sigma_n \) as defined in (5.7) is an injective field homomorphism, and hence \( K \cong \sigma_n(K) \) for each \( 1 \leq n \leq d \). Prove also that

\[
\mathbb{Q} = \{ \beta \in K : \sigma_n(\beta) = \beta \ \forall \ 1 \leq n \leq d \}.
\]

Problem 5.15. Let \( K \) be a number field of degree \( d \) so that \( K = \sigma_n(K) \) for each \( 1 \leq n \leq d \), where \( \sigma_1, \ldots, \sigma_n \) are embeddings of \( K \) into \( \mathbb{C} \). Prove that the set

\[
G := \{ \sigma_1, \ldots, \sigma_d \}
\]

is a group under the operation of function composition.

Problem 5.16. Let \( R \) be an integral domain and

\[
R^\times = \{ u \in R : \exists v \in R \text{ such that } uv = 1 \}
\]

the set of units in \( R \). Prove that \( R^\times \) is an abelian group under multiplication.

Problem 5.17. Prove that the group of units of the ring \( \mathbb{Z} \) is \( \{ \pm 1 \} \) and an element \( x \in \mathbb{Z} \) is an irreducible if and only if it is a prime.

Problem 5.18. Prove that the elements 2, 3, \( 1 \pm \sqrt{-5} \) are all irreducible in \( \mathcal{O}_{\mathbb{Q}(\sqrt{-5})} \), and 2, 3 do not divide \( 1 + \sqrt{-5} \) or \( 1 - \sqrt{-5} \).

Problem 5.19. Let \( \alpha_1, \ldots, \alpha_d \) and \( \beta_1, \ldots, \beta_d \) be two \( \mathbb{Q} \)-bases for the number field \( K \). Let \( C \) be the rational change of basis matrix from the \( \alpha \) to the \( \beta \) basis. Prove that

\[
\Delta(\beta_1, \ldots, \beta_d) = \det(C)^2 \Delta(\alpha_1, \ldots, \alpha_d).
\]

Problem 5.20. Let \( \sigma_1, \ldots, \sigma_d \) be embeddings of a number field \( K \), and let \( \alpha \in K \) be such that

\[
\sigma_j(\alpha) = \alpha \ \forall \ 1 \leq j \leq d.
\]

Prove that \( \alpha \in \mathbb{Q} \).

Problem 5.21. Let \( \alpha_1, \ldots, \alpha_d \) and \( \beta_1, \ldots, \beta_d \) be two integral bases for a number field \( K \). Prove that there exists a change of basis matrix \( A \in \text{GL}_d(\mathbb{Z}) \) between them.

Problem 5.22. Let \( R \) be a commutative ring and

\[
I_1 \subseteq I_2 \subseteq \ldots
\]

an ascending chain of ideals in \( R \). Prove that \( I = \bigcup_{k=1}^{\infty} I_k \) is also an ideal in \( R \).

Problem 5.23. Let \( K \) be a number field, \( \alpha, \beta \in K \) and \( a, b, c \in \mathbb{Q} \). Prove that

\[
N_K(c\alpha\beta) = c \cdot N_K(\alpha)N_K(\beta), \quad Tr_K(a\alpha + b\beta) = aTr_K(\alpha) + bTr_K(\beta).
\]
5.9. PROBLEMS

Problem 5.24. Let \( m \) be a positive integer and \( G \) be a finitely generated abelian group so that every element of \( G \) has finite order dividing \( m \). Prove that \( G \) is finite.

Problem 5.25. For an ideal \( I \subseteq \mathcal{O}_K \), let \( I' \) be as in (5.9). Prove that \( \mathcal{O}_K' = \mathcal{O}_K \).

Problem 5.26. Prove that for each \( B \in \mathfrak{F}_K \), \( B' \) is again a fractional ideal, i.e. \( B' \in \mathfrak{F}_K \) for every \( B \in \mathfrak{F}_K \).

Problem 5.27. Complete the proof of Lemma 5.6.1 by showing that if \( \mathcal{N}_K(IJ) = \mathcal{N}_K(I)\mathcal{N}_K(J) \) when \( J \) is a prime ideal, then this is true for all ideals \( J \).

Problem 5.28. Let \( I \subseteq \mathcal{O}_K \) be an ideal in the ring of integers \( \mathcal{O}_K \) of a number field \( K \), and \( P \subset \mathcal{O}_K \) a prime ideal. Define a map \( \phi: \mathcal{O}_K/IP \to \mathcal{O}_K/I \), given by \( \phi(x + IP) = x + I \). Prove that this is a surjective ring homomorphism.

Problem 5.29. Let \( R \) be a principal ideal domain. Prove that every \( \alpha \in R \) has a unique factorization into irreducibles.

Problem 5.30. Prove that \( \det(\Sigma(M)) \) in (5.13) of Lemma 5.7.1 cannot be equal to 0.

Problem 5.31. Prove that the set \( \mathfrak{P}_K \) of all principal fractional ideals is a subgroup of the group \( \mathfrak{F}_K \) of all fractional ideals in a number field \( K \).

Problem 5.32. Prove that two fractional ideals \( B_1, B_2 \in \mathfrak{F}_K \) are in the same ideal class, denoted by \( B_1 \sim B_2 \), if and only if there exists some \( a, b \in \mathcal{O}_K \) such that \( \langle a \rangle B_1 = \langle b \rangle B_2 \). Prove that this is an equivalence relation on the group \( \mathfrak{F}_K \).

Problem 5.33. Prove that every ideal class in \( \text{Cl}(K) \) contains an ideal \( I \subseteq \mathcal{O}_K \).
Appendices
Some properties of abelian groups

Here we briefly discuss some properties of abelian groups, in particular outlining a proof of the fact that any subgroup of a finitely generated abelian group is finitely generated. Throughout this section, we will mostly deal with a finitely generated abelian group \( G \), written additively with \( 0 \) denoting the identity element and \( nx \), for \( n \in \mathbb{Z} \) and \( x \in G \), denoting the \( n \)-th power of the element \( x \). A collection of elements \( x_1, \ldots, x_k \) in an abelian group \( G \) is called linearly independent if whenever

\[ n_1x_1 + \cdots + n_kx_k = 0 \]

for some \( n_1, \ldots, n_k \in \mathbb{Z} \), then \( n_1 = \cdots = n_k = 0 \). A linearly independent generating set for an abelian group \( G \) is called a basis. An abelian group \( G \) is called free if it has a basis. Hence free abelian groups are precisely lattices, and the most common example of a finitely generated free abelian group is \( \mathbb{Z}^k \), \( k \in \mathbb{Z}_{>0} \). In fact, it turns out that \( \mathbb{Z}^k \) is the only example of a finitely generated free abelian group, up to isomorphism.

**Lemma A.1.** Let \( G \) be a finitely generated free abelian group. Then \( G \cong \mathbb{Z}^k \) for some \( k \in \mathbb{Z}_{>0} \).

**Proof.** Let \( x_1, \ldots, x_k \) be a basis for \( G \), then

\[ G = \left\{ \sum_{i=1}^{k} n_ix_i : n_1, \ldots, n_k \in \mathbb{Z} \right\}. \]

Define a map \( \varphi : G \to \mathbb{Z}^k \), given by

\[ \varphi \left( \sum_{i=1}^{k} n_ix_i \right) = \sum_{i=1}^{k} n_ie_i. \]

We leave it to the reader to check that this is a group isomorphism. \( \square \)

**Corollary A.2.** Let \( G \) be a finitely generated free abelian group. Then every basis in \( G \) has the same cardinality. This common cardinality is called the rank of \( G \).

**Proof.** Let \( x_1, \ldots, x_k \) and \( y_1, \ldots, y_m \) be two different bases for \( G \). Then by the argument in the proof of Lemma A.1, \( G \cong \mathbb{Z}^k \) and \( G \cong \mathbb{Z}^m \). Now Problem A.2 implies that \( \mathbb{Z}^k \not\cong \mathbb{Z}^m \) unless \( k = m \). Recall that isomorphism is an equivalence relation on groups. Thus, since \( G \cong \mathbb{Z}^k \) and \( G \cong \mathbb{Z}^m \), we must have \( \mathbb{Z}^k \cong \mathbb{Z}^m \). Hence \( k = m \). \( \square \)

If \( H \) is a subgroup of a finitely generated free abelian group \( G \) of rank \( k \), then \( H \) is also free abelian of rank \( \leq k \); it is simply a sublattice of a lattice \( G \) of smaller rank. A standard proof is along the lines of linear algebra, using Smith normal
form for matrices, which constructs a basis for a subgroup starting with a basis for the group: it is very much along the lines of arguments given in Section 1.2.

We now recall some additional basic algebraic notation without proofs. We refer the reader to [DF03] for details. If \( G \) is an abelian group and \( H \) is a subgroup of \( G \), then a coset of \( H \) in \( G \) is a set \( x + H \) where \( x \in G \). The group \( G \) can be represented as a disjoint union of all cosets of \( H \) in \( G \). We write \( G/H \) for the set of such cosets, which is a group under the operation of addition of cosets:

\[
(x + H) + (y + H) = (x + y) + H.
\]

\( G/H \) is called the quotient group of \( G \) modulo \( H \). The identity element in this group is the trivial coset \( 0 + H = H = x + H \) for every \( x \in H \), and inverse of \( y + H \) is \(-y + H \) for every \( y \in G \). The order of \( G/H \), i.e. its cardinality as a set (could be infinite) is called the index of \( H \) in \( G \), and denoted by \(|G:H|\). Suppose that \( G \) and \( E \) are two abelian groups and \( \varphi : G \to E \) is a group homomorphism between them. Recall that \( \text{Ker}(\varphi) \) is a subgroup of \( G \) and \( \varphi(G) \) is a subgroup of \( E \). The First Isomorphism Theorem states that

\[
(A.1) \quad G/\text{Ker}(\varphi) \cong \varphi(G).
\]

Finally, notice that a finitely generated group can only be isomorphic to another finitely generated group. We are now ready for the main result of this section.

**Theorem A.3.** Let \( G \) be a finitely generated abelian group, and let \( H \) be a subgroup of \( G \). Then \( H \) is finitely generated.

**Proof.** Let us assume \( G \) is additively written. Let \( x_1, \ldots, x_k \) be a generating set for \( G \), then every element \( y \in G \) is expressible as

\[
y = \sum_{i=1}^{k} n_i x_i
\]

for some \( n_1, \ldots, n_k \in \mathbb{Z} \). Define a map \( \varphi : \mathbb{Z}^k \to G \), given by

\[
\varphi \left( \sum_{i=1}^{k} n_i e_i \right) = \sum_{i=1}^{k} n_i x_i.
\]

We leave it to the reader to check that this is a group homomorphism. Let \( K = \text{Ker}(\varphi) \), then \( K \) is a subgroup of \( \mathbb{Z}^k \), hence it is free abelian of rank \( \ell \leq k \). Now \( H \) be a subgroup of \( G \), then there exists a subgroup \( M \) of \( \mathbb{Z}^k \) such that \( \varphi(M) = H \); in other words, \( M \) is the pre-image of \( H \) in \( \mathbb{Z}^k \) under \( \varphi \). Then \( M \) is also free abelian of rank \( m \leq k \). Furthermore, \( M \) contains \( K \); indeed, for every \( x \in K \), \( \varphi(x) = 0 \in H \), hence \( x \in M \). Therefore \( \ell \leq m \), and by (A.1),

\[
H \cong M/K,
\]

hence we only need to show that \( M/K \) is finitely generated.

By Lemma A.1 we know that \( M \cong \mathbb{Z}^m \) and \( K \cong \mathbb{Z}^\ell \). By viewing vectors in \( \mathbb{Z}^\ell \) as \( m \)-tuples with last \( m-\ell \) coordinates equal to 0, we can think of \( \mathbb{Z}^\ell \) being contained in \( \mathbb{Z}^m \). Hence we only need to show that \( \mathbb{Z}^m/\mathbb{Z}^\ell \) is finitely generated. If \( m = \ell \), then \( \mathbb{Z}^m = \mathbb{Z}^\ell \) and so \( \mathbb{Z}^m/\mathbb{Z}^\ell \cong \{0\} \), the trivial group. Then assume that \( m > \ell \). Considering the standard basis \( e_1, \ldots, e_m \) for \( \mathbb{Z}^m \), we can view \( e_1, \ldots, e_\ell \) as the standard basis for \( \mathbb{Z}^\ell \) under its embedding into \( \mathbb{Z}^m \). Then \( \mathbb{Z}^m/\mathbb{Z}^\ell \) is isomorphic to \( \mathbb{Z}^{m-\ell} \) via the map sending an element \( \sum_{i=1}^{m} n_i e_i + \mathbb{Z}^\ell \) in \( \mathbb{Z}^m/\mathbb{Z}^\ell \) to \( \sum_{i=m-\ell+1}^{m} n_i e_i \).
Problems

**Problem A.1.** Suppose that $G$ is a free abelian group. Prove that the following property holds: whenever $nx = 0$ for some $n \in \mathbb{Z}$ and $x \in G$, then either $n = 0$ or $x = 0$.

**Problem A.2.** Suppose that $1 \leq k < m$. Prove that free abelian groups $\mathbb{Z}^k$ and $\mathbb{Z}^m$ are not isomorphic.
APPENDIX B

Maximum Modulus Principle and Fundamental Theorem of Algebra

Our main goal here is to prove the Fundamental Theorem of Algebra. For this, we will use the Maximum Modulus Principle. We first need some basic notation from complex analysis. A region in \( \mathbb{C} \) is a subset \( R \) of \( \mathbb{C} \), which is open and connected. A function \( f(z) \) on a region \( R \) is called analytic if for any \( z_0 \in R \),

\[
f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n,
\]

where \( a_n \in \mathbb{C} \) for every \( n \geq 0 \) and the series is convergent to \( f(z) \) in an an open neighborhood of \( z_0 \). It is a well-known fact that every holomorphic (i.e., complex-differentiable) function is analytic, and vice versa.

**Theorem B.1** (Maximum Modulus Principle). Suppose \( f(z) \) is a non-constant analytic function in a region \( R \). Then the real-valued function \( |f(z)| \) does not attain its maximum in \( R \). In other words, if for some \( z_0 \in R \), \( |f(z)| \leq |f(z_0)| \) for all points \( z \in R \), then \( f(z) \) is constant on \( R \).

A proof of this theorem can be found in any book on complex analysis, for instance [Rud87]. Here is an immediate consequence of Theorem B.1, which is very useful in applications.

**Corollary B.2.** Let

\[
D_r = \{ z \in \mathbb{C} : |z| \leq r \}
\]

be the closed disk of radius \( r \) and let \( f(z) \) be a continuous function on \( D_r \), which is analytic on the open disk

\[
D^o_r = \{ z \in \mathbb{C} : |z| < r \}.
\]

Then \( f(z) \) assumes its maximum value on \( D_r \) on its boundary

\[
\partial D_r = \{ z \in \mathbb{C} : |z| = r \} = D_r \setminus D^o_r.
\]

**Proof.** Since \( f(z) \) is continuous and \( D_r \) is closed and bounded, \( f(z) \) must have a maximum on \( D_r \). On the other hand, since the open disk \( D^o_r \) is a region in \( \mathbb{C} \), by Theorem B.1 \( f(z) \) cannot have a maximum on \( D^o_r \). Thus it must be assumed on the boundary. \(\square\)

We will now derive an important consequence of this fundamental principle.

**Theorem B.3.** [Fundamental Theorem of Algebra] Any polynomial \( p(x) \in \mathbb{C}[x] \) of degree \( n \) has precisely \( n \) roots in \( \mathbb{C} \), counted with multiplicity. In other words, the field of complex numbers \( \mathbb{C} \) is algebraically closed.
Proof. Notice that it is sufficient to prove that any polynomial $p(x)$ of degree $n \geq 1$ has at least one root in $\mathbb{C}$. Suppose not, say $p(x) \in \mathbb{C}[x]$ of degree $n \geq 1$ has no complex roots. This means that $1/p(x)$ is an analytic (holomorphic) function. Notice that $1/p(x)$ tends to zero as $|x|$ tends to infinity. This means that for any $\alpha \in \mathbb{C}$ there exists an $r \in \mathbb{R}$ such that

$$1/|p(x)| < 1/|p(\alpha)|$$

for all $x \in \mathbb{C}$ with $|x| \geq r$. Now pick $r$ large enough so that $|\alpha| < r$, and let $D_r$ be the closed disk of radius $r$, as in Corollary B.2 above. Then $\alpha \in D_r$ and, since $1/|p(x)|$ is continuous, it assumes its maximum on $D_r$, specifically on its boundary, by Corollary B.2. Then there exists $\beta \in \partial D_r$ such that

$$1/|p(x)| \leq 1/|p(\beta)| \quad \forall x \in D_r.$$

Now pick $t > r$ and $D_t^\circ$ be the open disk of radius $t$. Then $D_r \subsetneq D_t^\circ$, and for all $x \in D_t^\circ \setminus D_r$,

$$1/|p(x)| < 1/|p(\alpha)| \leq 1/|p(\beta)|.$$

Hence $1/|p(x)|$ assumes its maximum on $D_t^\circ$ at $x = \beta$. Since $1/p(x)$ is not a constant function (degree of $p(x)$ is $> 0$) and $D_t^\circ$ is a region (it is open and connected), this violates the Maximum Modulus Principle. Hence $p(x)$ must have a zero in $\mathbb{C}$. \qed
Bibliography


